データセット説明文からの数値メタデータ自動抽出

嘉本名晋 [1]　田村晃裕 [1]　桂井麻里衣 [1]
[1] 同志社大学大学院 理工学研究科
{kamoto24,katsurai}@mm.doshisha.ac.jp, aktamura@mail.doshisha.ac.jp

概要

機械学習研究において適切なデータセットの発見は重要な課題であるが，データセット説明文の多くは非構造化テキストであり，数値による検索が困難である．本研究では，データセット文書から数値情報を抽出し構造化メタデータとして整理するタスクを新たに提案する．まず，論文とリポジトリからデータセットに関する説明文書 3,926 件を抽出し，数値エンティティとその文脈に関するアノテーション済みコーパスを構築した．次に，数値情報と数値情報の補足情報（非数値情報）を二段階に分けて抽出するようモデルを訓練した．実験では，数値エンティティに対して F1 スコア 72.0%，非数値エンティティに対して F1 スコア 63.4%を達成した．

1　はじめに

機械学習モデルの訓練と評価においてデータセットは不可欠な要素であり，様々な分野において数千を超えるデータセットが公開されている．これらのデータセットを選択する際には，データサイズや収集方法，アノテーション方式など複数の要因を考慮する必要がある．しかし，データセット情報は論文やリポジトリページに散在しており，その多くは非構造化テキストである．数値情報が統一的な形式で記述されていないため，研究者は実験要件を満たすデータセットを見つけることが困難な場合がある．

この問題に対する一つのアプローチとして，非構造化文書から数値情報を自動抽出することが挙げられる（図 1）．このようなシステム構築には，ラベル体系の設計，対象領域におけるアノテーション済みコーパスの構築，エンティティ抽出モデルの学習が必要となる．しかしデータセット文書を対象とした一貫性のあるアノテーション済みコーパスや標準的なベンチマークは存在せず，モデル開発と公平な比較評価が制限されている．

本論文ではこれらの課題に対処するため，

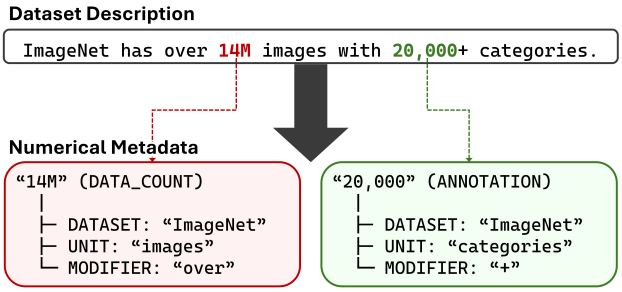

図 1　数値メタデータ抽出の概要

データセット文書からの数値情報抽出コーパス NIED（Numerical Information Extraction from Dataset descriptions）を提案する．NIED は学術論文，リポジトリ説明文から収集した 3,926 件のアノテーション付き文書で構成される．実際のデータセット作成手順を反映した二層のラベル体系として，データ量やアノテーションの種類数などの数値エンティティと，その意味的文脈を提供する非数値エンティティを区別することで，データセットの様々な特性を表すメタデータを生成する．

七つの事前学習済み言語モデルを用いた二段階抽出手法によるベースライン評価では，最良モデルが数値エンティティに対して F1 スコア 72.0%，非数値エンティティに対して F1 スコア 63.4%を達成し，データセットメタデータ整理のための自動抽出の実現可能性を示した．本リソース[1)]はデータセットの定量的特性を活用した検索・比較ツールの開発を促進することを目的とする．

2　関連研究

数値情報抽出は様々な領域で必要とされているものの，これまで十分な研究が行われてこなかった [1]．先駆的な取り組みとして MeasEval 共有タスク [2] が科学論文からのカウント，測定値，関連文脈の抽出に関するベンチマークを確立した．しかし，MeasEval のラベル体系は化学や生物医学における

1）https://github.com/mm-doshisha/NIED

表 1　ラベル定義とエンティティ分布

数値ラベル		
ラベル	**定義**	**数**
DATA_COUNT	データ数に関する数値．全体サイズ，訓練・検証・評価データ数など	1,244
ANNOTATION	アノテーションの種類数．クラス数，カテゴリ数，タグ数など	259
COLLECTION	データ収集に関する数値．被験者数，収集源数，収集条件数など	395
DOMAIN	対象ドメインの数．言語数，分野数，テーマ数など	53
CONTRIBUTOR	データセット作成に関与した人数．アノテーター数，評価者数など	49
非数値ラベル		
ラベル	**定義**	**数**
UNIT	数値が何を数えているかを示す基本単位	1,871
EXTENDED_UNIT	UNIT に修飾語を加えた詳細な単位表現	1,860
QUALIFIER	数値の意味を明確化する補足情報．データ特性，収集方法，出所，用途など	1,201
MODIFIER	数値の精度や範囲の程度を示す表現．approximately，at least，average など	197
DENOMINATOR	数値が表す比率や密度の基準となる単位	116
SPLIT	数値が属するデータ分割を示す用語．training，validation，test など	200
DATASET	数値に関連するデータの格納・提供元．データセット名，コーパス名など	526

物理量の識別を目的として設計されており，データセット文書への適用可能性は限定的である．

近年，学術文書を対象とした領域特化型の抽出タスクに関する研究も行われている．例えば Saier ら [3] は機械学習論文における実験設定パラメータの抽出に焦点を当て，Alyafeai ら [4] はスキーマ駆動アプローチを用いた科学論文からの構造化メタデータ抽出の実現可能性を示した．しかし，実験パラメータや物理測定値とは異なり，データセット構築情報には以下の二つの特徴的性質がある．第一に，数値はデータセット作成手順において異なる役割，すなわちデータ量，アノテーションの種類数，収集条件などを果たすため意味的カテゴリ化が必要である．第二に，数値エンティティ単独では意味のある解釈に不十分であり，その正確な意味を決定する文脈要素を抽出する必要がある．そのため本研究ではデータセット文書の構造に適したラベル体系の設計から開始した．

3　コーパス構築

3.1　データ収集

データセットのモダリティと領域の多様性を確保するため，以下 4 つの情報源からデータセット文書を収集した．

Papers with Code (PwC)[2]：画像，テキスト，動画，音声の 4 モダリティから各 50 件の説明文を収集し計 200 件を得た．

Figshare[3]：コンピュータサイエンス以外の領域の多様性確保のため，同サイトで定義されている 22 トップレベルカテゴリから各 20 件をサンプリングし，多様な学術分野にわたる 440 件の説明文を得た．

NeurIPS Datasets and Benchmarks track (2021–2024)：arXiv と OpenReview プラットフォームから，このトラックで採択された 40 件の論文を収集した．

arXiv：研究論文における文書スタイルの網羅性確保のため，Papers with Code から直接参照されている 40 件の arXiv 論文を追加収集した．

全論文は CC-BY ライセンスのものに限定し，GROBID[4] を用いてテキスト抽出を行った．図表，引用，一般的なセクション（Introduction，Related Work，Conclusion）は除外した．spaCy[5] を用いて品詞タグを付与した後，少なくとも 1 つの NUM タグを持つトークンを含む文のみを有効なデータセット文書として保持した．

3.2　アノテーションと品質評価

先行研究 [2] を参考に本タスクのアノテーションガイドラインを作成し，第一著者が全文書に対し

2）https://paperswithcode.com/

3）https://figshare.com/
4）https://github.com/kermitt2/grobid
5）https://spacy.io/

表 2　コーパス統計

項目	値
総文書数	3,926
総文数	16,727
数値エンティティスパン数	2,000
非数値エンティティスパン数	5,971

て手動でアノテーションを行った．表 1 に 12 ラベルの定義と分布を，表 2 にコーパスの統計情報を示す．

品質評価のため，第一著者以外の 3 名のアノテーターがコーパスの 10%に相当するサブセットに対してアノテーションを行った．第一著者を含む 4 名のアノテーター間の全ペアについて一致度を測定し，その平均を算出した．一致度の評価にはエンティティスパンの一致を評価するための pairwise F1 score を用いた．数値ラベルの Micro F1 スコアは 80.5%，非数値ラベルの Micro F1 スコアは完全一致で 75.6%，境界一致で 79.7%を達成した．詳細な評価結果は付録に示す．

4　二段階抽出手法

対象タスクでは複数の非数値エンティティがトークン境界を共有する場合がある．例えば「1,000 training samples」というフレーズでは，training を SPLIT として，training samples を EXTENDED_UNIT としてアノテーションするため，これらのエンティティは training というトークンを共有し境界が重複する．本手法は抽出を逐次的なステップに分解することでこの問題に対処する．まず数値エンティティを抽出し，次にそれらをアンカーとして用いて非数値エンティティを抽出する．この設計は MeasEval で高性能を示したカスケード方式 [5] から着想を得ている．

4.1　ステップ 1：数値エンティティ抽出

このステップでは数値エンティティを識別・分類する標準的な固有表現抽出を行う．具体的には，アノテーション付きデータセットで事前学習済み言語モデルをファインチューニングし，5 つの数値ラベル，すなわち DATA_COUNT，ANNOTATION，COLLECTION，DOMAIN，CONTRIBUTOR を予測する．

4.2　ステップ 2：非数値エンティティ抽出

ステップ 1 で検出された各数値エンティティに対し，このステップではその文脈を提供する非数値エンティティを抽出する．検出された数値エンティティを特殊トークン [NUM] と [/NUM] で囲み，七つの非数値ラベルそれぞれに対して個別の二値分類モデルを訓練する．複数の数値エンティティを含む場合，各エンティティに対して個別の入力インスタンスを生成し，対象の数値エンティティのみをマークする．最終的に全二値分類モデルからの最終予測を集約し，数値および非数値エンティティとそれらの対応関係を示す包括的なアノテーションを生成する．

5　実験

5.1　実験設定

RoBERTa [6]，BERT [7]，DistilBERT [8]，ALBERT [9]，XLNet [10]，DeBERTa [11]，SciBERT [12] の七つの事前学習済みモデルを用いて二段階手法を実行した．データセットを訓練セットとテストセットに 4:1 の比率で分割した後，各モデルを数値エンティティ抽出と七つの非数値ラベルそれぞれに対して個別にファインチューニングした[6]．比較手法として条件付き確率場（CRF）と BiLSTM-CRF [13] も訓練した．

評価にはスパンベースの評価指標を用い，完全一致と境界一致の両スキームで適合率，再現率，F1 スコアを算出した．完全一致はスパン境界が完全に一致した場合のみ，境界一致は部分的にでも重複がある場合にエンティティを正解とする．なお，非数値エンティティは数値エンティティレベルで評価し，複数の数値エンティティと対応している場合は各対応を個別にカウントした．

5.2　結果と分析

表 3 より，BERT ベースモデルがデータセット文書からの数値情報抽出において CRF ベース手法を大幅に上回ることが分かる．XLNet が最高性能を達成し，完全一致評価において数値エンティティに対して F1 スコア 72.0%，非数値エンティティに対して F1 スコア 63.4%を記録した．

6）ファインチューニングには spacy init fill-config で自動生成された設定値を使用した．詳細な設定は GitHub リポジトリで公開している．

表 3　数値および非数値エンティティ抽出におけるモデル性能比較

モデル	数値						非数値					
	完全一致			境界一致			完全一致			境界一致		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ALBERT	0.778	0.639	0.702	0.795	0.653	0.717	0.674	0.563	0.613	0.729	0.608	0.663
BERT	0.726	**0.684**	0.704	0.741	**0.698**	0.719	0.622	**0.598**	0.610	0.680	**0.654**	0.666
DeBERTa	0.781	0.658	0.714	0.801	0.675	0.733	0.677	0.577	0.623	0.729	0.622	0.671
DistilBERT	0.771	0.642	0.700	0.791	0.658	0.719	**0.688**	0.571	0.624	**0.743**	0.617	0.674
RoBERTa	**0.804**	0.633	0.709	**0.826**	0.650	0.727	0.679	0.556	0.611	0.739	0.606	0.666
SciBERT	0.779	0.670	**0.720**	0.798	0.686	**0.738**	0.664	0.581	0.620	0.731	0.639	0.682
XLNet	0.779	0.670	**0.720**	0.795	0.684	0.735	0.684	0.590	**0.634**	0.740	0.638	**0.685**
CRF	0.735	0.451	0.559	0.744	0.457	0.566	0.574	0.295	0.390	0.623	0.320	0.422
BiLSTM-CRF	0.413	0.417	0.415	0.443	0.448	0.446	0.208	0.213	0.210	0.302	0.310	0.306

表 4　最良モデルのラベル別性能

ラベル	F1 スコア	
	完全一致	境界一致
数値ラベル (XLNet)		
ラベル非依存検出	**0.762**	**0.777**
ANNOTATION	0.500	0.500
COLLECTION	0.656	0.656
CONTRIBUTOR	0.769	0.769
DATA_COUNT	0.771	0.794
DOMAIN	0.737	0.737
Macro Average	**0.687**	**0.691**
Micro Average	**0.720**	**0.735**
非数値ラベル (XLNet)		
DATASET	0.622	0.693
DENOMINATOR	0.409	0.409
EXTENDED_UNIT	0.671	0.720
MODIFIER	0.700	0.700
QUALIFIER	0.551	0.652
SPLIT	0.424	0.471
UNIT	0.688	0.713
Macro Average	**0.581**	**0.623**
Micro Average	**0.634**	**0.685**

表 4 に XLNet の詳細な性能内訳を示す．分類を伴わない数値エンティティ検出は完全一致で F1 スコア 76.2%，境界一致で F1 スコア 77.7%を達成したが，特定のラベルへの分類を伴う場合，Micro F1 スコアは完全一致で 72.0%，境界一致で 73.5%に低下した．この結果は，数値の位置特定は適切なラベルの割り当てよりも容易であることを示唆している．

性能はラベルの出現頻度と複雑さによって変動した．最も出現頻度の高い DATA_COUNT が最良の性能を示し，完全一致で F1 スコア 77.1%を記録した．一方，ANNOTATION は分類が最も困難であり，完全一致で F1 スコア 50.0%にとどまった．これは訓練データが少ないことに加え，ANNOTATION ラベルが示す数値と DATA_COUNT など他のラベルが示す数値を区別することが難しいためと考えられる．同様に，非数値ラベルの中では UNIT と EXTENDED_UNIT が良好な性能を示し，より高い出現頻度と明確な語彙パターンの恩恵を受けた．

6　おわりに

本論文では，データセット文書から構造化数値情報を抽出する課題に取り組み，NIED コーパスの構築と二層ラベル体系の提案という二つの主要な貢献を行った．実験により，事前学習済み言語モデルを用いた数値および非数値エンティティ両方の自動抽出の実現可能性を確認した．今後の課題として，研究基盤への応用を探求し，データセットの発見可能性とメタデータ標準化を強化するための既存学術データベースとの統合を検討する．

謝辞

本研究は JSPS 科研費 JP25K03419 の助成を受けた．また，本研究は東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

参考文献

[1] Jan Göpfert, Patrick Kuckertz, J Weinand, Leander Kotzur, and D Stolten. Measurement extraction with natural language processing: A review. In *EMNLP*, pp. 2191–2215, 2022.

[2] Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts. In *Proc. SemEval-2021*, 2021.

[3] Tarek Saier, Mayumi Ohta, Takuto Asakura, and Michael Färber. HyperPIE: Hyperparameter information extraction from scientific publications. In *ECIR 2024*, pp. 254–269, 2024.

[4] Zaid Alyafeai, Maged S Al-Shaibani, and Bernard Ghanem. MOLE: Metadata extraction and validation in scientific papers using LLMs. *arXiv preprint arXiv:2505.19800*, 2025.

[5] Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. LIORI at SemEval-2021 task 8: Ask Transformer for measurements. In *Proc. SemEval-2021*, pp. 1249–1254, 2021.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[9] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 2019.

[11] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

表 5　数値ラベルのアノテーション品質評価

ラベル	F1 スコア
ラベル非依存検出	**0.886**
ANNOTATION	0.642
COLLECTION	0.684
CONTRIBUTOR	0.900
DATA_COUNT	0.850
DOMAIN	0.583
Micro Average	**0.805**
Macro Average	**0.732**

表 6　非数値ラベルのアノテーション品質評価

ラベル	F1 スコア	
	完全一致	境界一致
DATASET	0.663	0.719
DENOMINATOR	0.609	0.671
EXTENDED_UNIT	0.839	0.859
MODIFIER	0.723	0.730
QUALIFIER	0.525	0.660
SPLIT	0.864	0.864
UNIT	0.840	0.855
Micro Average	**0.756**	**0.797**
Macro Average	**0.723**	**0.766**

付録

付録 A：アノテーション品質評価の詳細

アノテーション品質評価と人間性能のベースライン確立のため，同志社大学の学部生 3 名を採用しコーパスの一部を評価した．練習セッション完了後，Figshare 説明文 44 件，Papers with Code 説明文 20 件，Papers with Code 論文 4 件，NeurIPS 論文 4 件からなる評価用サブセットに対してアノテーションを行った．

第一著者を含む 4 名のアノテーター間の全ペアについて一致度を測定し，その平均を算出した．一致度の評価にはエンティティスパンの一致を評価するための pairwise F1 score を用いた．またラベルに関係なく数値エンティティの識別に関する一致度を測定するラベル非依存の数値エンティティ検出も評価した．

表 5 と表 6 に数値および非数値ラベルに対する品質評価結果を示す．数値ラベルでは完全一致と境界一致で同じ値が得られたため表 5 には 1 つの結果のみを示す．結果は両タイプのラベルにわたって妥当な一貫性を示している．

表 7　ゴールド数値エンティティ条件下での非数値ラベル抽出性能

モデル	完全一致		境界一致	
	Macro	Micro	Macro	Micro
ALBERT	0.747	0.820	0.786	0.868
BERT	0.730	0.813	0.788	0.868
DeBERTa	**0.754**	0.820	0.794	0.863
DistilBERT	0.749	**0.820**	0.798	0.868
RoBERTa	0.748	0.813	0.798	0.862
SciBERT	0.746	0.810	0.800	0.866
XLNet	0.749	0.820	**0.807**	**0.874**

付録 B：ゴールド数値エンティティ条件下での非数値ラベル抽出性能

提案手法の二段階アプローチにおける非数値エンティティ抽出の性能上限を評価するため，ステップ 1 で抽出された数値エンティティの代わりに正解の数値エンティティを与えた条件下でステップ 2 の性能を測定した．この実験設定では数値エンティティの抽出誤差を排除し，非数値エンティティ抽出モデルの純粋な性能を評価できる．

表 7 にゴールド数値エンティティ条件下での各モデルの非数値ラベル抽出性能を示す．全モデルが通常条件と比較して大幅な性能向上を示し，Micro F1 スコアで約 18%から 20%の改善が見られた．最良モデルは境界一致評価において XLNet であり，Micro F1 スコア 87.4%を達成した．

ゴールド条件下でも完全一致において Micro F1 スコアが 82%程度にとどまることから，非数値エンティティ抽出自体に改善の余地があることが分かる．興味深いことに，XLNet は完全一致では DistilBERT や DeBERTa に劣るものの境界一致では最良の性能を示しており，境界の柔軟性において優位性を持つことが示唆される．この結果は，ステップ 1 における数値エンティティ抽出の精度向上が全体の性能に大きく寄与する一方で，非数値エンティティ抽出モデル自体の改善も重要であることを示している．