テキストアナリティクスツールの説明文に含まれる設定キーの認識

仲田将斗 [1]　亀甲博貴 [2]　森信介 [2]
[1] 京都大学大学院 情報学研究科　[2] 京都大学 学術情報メディアセンター
nakata.masato.26m@st.kyoto-u.ac.jp　{kameko,forest}@i.kyoto-u.ac.jp

概要

テキストアナリティクスツールは，テキスト集合の性質を統計的に明らかにすることを目的として，人文学において広く用いられている．こうした分野において，テキストアナリティクス実験の再現性を保証するために，論文の説明文から分析手法やそのパラメータを正しく特定することが大事となる．我々はこれをいくつかのサブタスクへ分解し，そのうちの一つである設定キー認識タスクに焦点を当てた．ここでは設定キーごとの二値分類として定式化し，事前学習済みの言語モデルに基づくencoder-decoder モデルを提案する．

1　はじめに

人文学分野において，研究対象のテキスト集合を統計的手法によって分析するために，テキストアナリティクスツールが注目を集めている [1]．こうした手法は，単語の出現頻度による分析やクラスター分析，潜在的ディリクレ配分法（LDA）など多岐に渡るが，これらの手法を使いこなすには情報学の専門知識が必要となる．それは分析結果の論文を読み書きする際も例外ではなく，分析手法のパラメータに関して必要十分に説明文を記述する，あるいはその説明文を読み解き正しくパラメータを特定する，という作業は特に非専門家にとって苦労を要している．

こうした労力を軽減するために，森田ら [2] はテキストアナリティクスツールの操作ログと説明文の間の相互変換モデルを提案した．操作ログ（以下ログと表記）とはテキストアナリティクスの分析手法とそのパラメータの対のことを言い，ツールが直接扱えるような構造化データとなっている．森田らはログと説明文の正解対からなるデータセットを作成し，これを「原文と翻訳文」の対と見做すことで言語モデルによる変換モデルを実現した．

森田らは「ログからの説明文生成」と「説明文からのログ生成」の双方向の変換を試みたが，そのうちログ生成タスクにおいて以下の問題を抱えている：

- ログは構文解析できるようにフォーマットが定められているが，出力が正しいフォーマットに従っているとは限らない．
- 存在しないパラメータや表記の僅かに誤ったパラメータを出力することがある．

そこで我々はログを “生成” するのではなく “認識・抽出” するという立場で，実用的な精度まで引き上げることを目標とする．

まずログ生成タスクをより細かなサブタスクに分解し，認識や抽出のタスクに帰着させる．分解の詳細については 2 節で述べる．続いてサブタスクの一つである**設定キー認識**，つまり分析手法のパラメータのうち「どのパラメータが明示的にセットされているか（デフォルト値でないか）」を特定するタスクを取り上げ，それを解くためのモデルを 3 節で提案する．最後に，このモデルの精度評価について 4 節で解説する．

2　タスク詳細

本節ではまずタスクに係る用語の定義をし，ログ生成タスクについて述べる．その後，ログ生成タスクのサブタスクを説明し，本研究での対象範囲を規定する．

2.1　用語の定義

一つのテキストアナリティクスツールは，**分析手法**と呼ばれる機能（*functions*）の集合 $\mathfrak{F}$ を持つ．分析手法は，ユーザが与えたテキスト集合を分析するための統計的手法または機械学習による手法であり，たとえばあるテキストアナリティクスツール

には

$$\text{共起ネットワーク}, \text{対応分析} \in \mathcal{F}$$

といった手法が存在する．

各分析手法 $f \in \mathcal{F}$ は固有の**設定キー**（*configuration key*）集合 $\mathcal{K}_f$ を持ち，それぞれの設定キー $k \in \mathcal{K}_f$ には設定可能な**設定値**（*configuration values*）の集合 $\mathcal{V}_{f,k}$ が対応する．これらの集合はツールの仕様によって事前に定められている．たとえばあるテキストアナリティクスツールの共起ネットワーク（$\mathsf{CN} \in \mathcal{F}$ と表記する）には

$$\mathsf{MinTF} := \text{語の最小出現数} \in \mathcal{K}_{\mathsf{CN}}$$

などの設定キーがあり，設定値として 0 以上の整数を指定できる：

$$\mathcal{V}_{\mathsf{CN},\mathsf{MinTF}} \subset \mathbb{Z}_{\geqslant 0}.$$

ところで，設定値というものは必ずしも明示的に指定されるわけではなく，デフォルトの値が暗黙的に用いられることも多い．そのため，分析手法 $f \in \mathcal{F}$ の**設定**を，設定キー集合のある部分集合 $K \subset \mathcal{K}_f$ に対する設定値の対応

$$\{(k, v_k) \mid k \in K\} \quad (v_k \in \mathcal{V}_{f,k})$$

として定義し，K に含まれない設定キーについてはデフォルト値であると見做す．このような定式化はタスクの本質には影響しないが，これによって後述の設定キー認識を単純なモデルで解くことができるようになる（3 節）．そこでは，説明文中には K に含まれる設定キーのみ現れるという仮定を置く．

最後に，テキストアナリティクスツールの**操作ログ**あるいは単に**ログ**とは，分析手法 $f \in \mathcal{F}$ とその設定 c の対 (f, c) を意味し，それを自然言語により論文等で説明したものを**説明文**と呼ぶ．

2.2 ログ生成タスクとその分解

このように定式化したとき，ログ生成タスクは「説明文 D が与えられたとき，条件付き確率 $p(L \mid D)$ が最大となるログ L を生成せよ」というタスクと言い換えることができる．このログ L を分析手法 F と設定 C という二つの確率変数によって書き直せば，

$$p(L \mid D) = p(F, C \mid D) = p(C \mid F, D) \cdot p(F \mid D)$$

という関係を得る．さらに設定を設定キーの集合 $K \subset \mathcal{K}_F$ とそれらの値対応 V に分けて概念的に

$$p(C \mid F, D) = p(V \mid K, F, D) \cdot p(K \mid F, D)$$

のような分解を考えることができる．

以上から示唆を得て，ログ生成タスクを次の三つのサブタスクへ分解する．

1. 分析手法分類 —— $p(F \mid D)$ に対応．説明文がどの分析手法に対しての記述なのかを判定する．
2. 設定キー認識 —— $p(K \mid F, D)$ に対応．説明文中に明示された設定キーを列挙する．
3. 設定値抽出 —— $p(V \mid K, F, D)$ に対応．説明文中からそれぞれの設定値を特定し，設定キーとの対応付けを構築する．

設定キー認識および設定値抽出に関しては図 1 も見よ．

2.3 本研究の範囲

上で見た三つのサブタスクのうち，本研究では 2. 設定キー認識を扱う．

3 提案モデル

設定キー認識（2.2 節）を解くためのモデルとして，設定キーごとの二値分類器を提案する．

より詳細に説明するために，設定

$$c = \{(k, v_k) \mid k \in K\}$$

を持つログ $\ell = (f, c)$ を考えよう．記法については 2.1 節を見よ．このログから作られた説明文 d があるとして，設定キー認識とは d から設定キー集合 $K \subset \mathcal{K}_f$ を推定する問題となる．（分析手法 f は予め推定できているとする．）そこで我々が目指すのは，各設定キー $k \in \mathcal{K}_f$ に対して二値分類器

$$\mathsf{Found}_k : \{\text{説明文}\} \to \{0, 1\}$$

を，与えられた説明文 d について

$$\{k \in \mathcal{K}_f \mid \mathsf{Found}_k(d) = 1\} = K$$

が成り立つように構築することである．つまり，$\mathsf{Found}_k(d) = 1$ は説明文の元となったログ ℓ 中に設定キー k が**明示的に**現れていることを意味し，$= 0$ は明示的には現れずデフォルト値が使われていることを意味する．たとえば図 1 の入力例 d の場合であれば，$\mathsf{Found}_{\text{上位}}(d) = 1$，$\mathsf{Found}_{\text{外部変数}}(d) = 0$ などとなる．

そしてこの分類器を，事前学習された言語モデルによる transformer encoder と，線形 decoder からなる確率モデルで実現する（図 2）．Transformer encoder

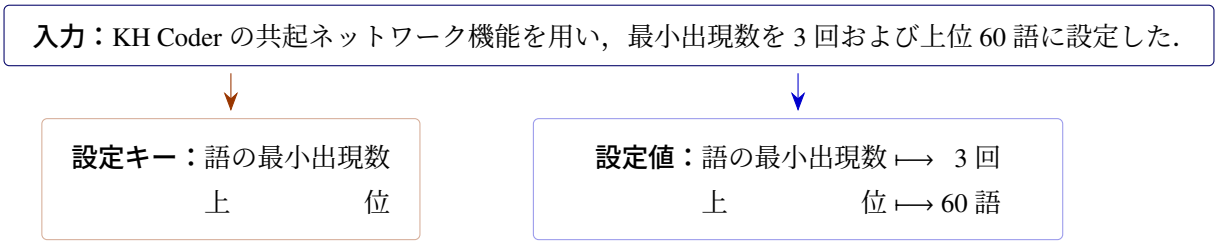

図 1 設定キー認識（左）および設定値抽出（右）．

には BERT [3] を用い，説明文を表現ベクトルの列に変換させる．線形 decoder 層はその表現ベクトルを受け取り，線形変換した後 活性化関数によって 0 以上 1 以下の実数を出力する．

比較として，transformer encoder に BERT の代わりに T5 [4]，GPT-2 [5]，RoBERTa [6] を使用したモデルも提案する．T5 などは transformer encoder-decoder ではあるが，単にある次元のベクトルを出力するニューラルネットワークと見做し，BERT の部分を置き換えた．

4 実験

2.2 節で述べたサブタスクへの分解，および 3 節のモデルを検証するための実験を行った．ベースラインのモデルとして，サブタスクへの分解をしない森田ら [2] によるモデルの実験も行った．

4.1 データセット

実験には森田らによって作成されたデータセットを用いた．このデータセットは，広く普及したテキストアナリティクスツールである KH Coder [7] を用いた実際の論文を集め，それらにログと説明文をアノテーションすることで作られた．共起ネットワークおよび対応分析の二つの分析手法について，合わせて約 250 エントリーを収録している．

4.2 実験設定

4.2.1 提案モデル

事前学習 それぞれのモデルの encoder には，日本語の大規模なコーパスで事前学習され，公開されているものを用いた．事前学習に使用したコーパスには Wikipedia，CC 100 [8, 9]，OSCAR [10, 11] がある（表 1）． CC 100 および OSCAR は多言語を含むコーパスであるが，そのうち日本語の部分に限定している．

モデル	Wikipedia	CC 100	OSCAR
BERT	○	○	-
T5	○	○	○
GPT-2	○	○	-
RoBERTa	○	○	-

表 1 Encoder に用いたモデルを事前学習したコーパス．○とマークされたコーパスが使用された．

メトリクス $\mathrm{Found}_k(d) = 1$ を陽性（positive），$= 0$ を陰性（negative）として，全体に占める真陽性と真陰性の割合

$$\mathrm{acc} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

をメトリクスとして採用した．

4.2.2 先行研究のモデル

森田らの提案した，T5 [4] によるマルチタスクモデルを用いた．T5 モデルは，我々のモデルと同じコーパスで事前学習されている．学習時（validation データ上での評価）のメトリクスとして，森田らは 4-gram に対する BLEU [12] を採用したが，データセットのサイズを考慮して代わりに 1-gram に対する BLEU を採用した．

テストデータ上での評価の際には，出力されたログから設定キーを抽出し，それを設定キー認識の結果とした．この結果をさらに設定キーごとに分け，それぞれの設定キーについて上記のメトリクス acc を計算した．なお，データセット中のログはすべて一定のフォーマットに従っているが，言語モデルの出力も同じフォーマットであるとは限らない．そこで，正しくないフォーマットの場合は設定キーが存在しない，つまり 2.1 節の記法で $K = \emptyset$ と見做して評価を行った．

4.2.3 共通の設定

交差検証 データセットのサイズを考慮し，5-fold 交差検証を実施した．

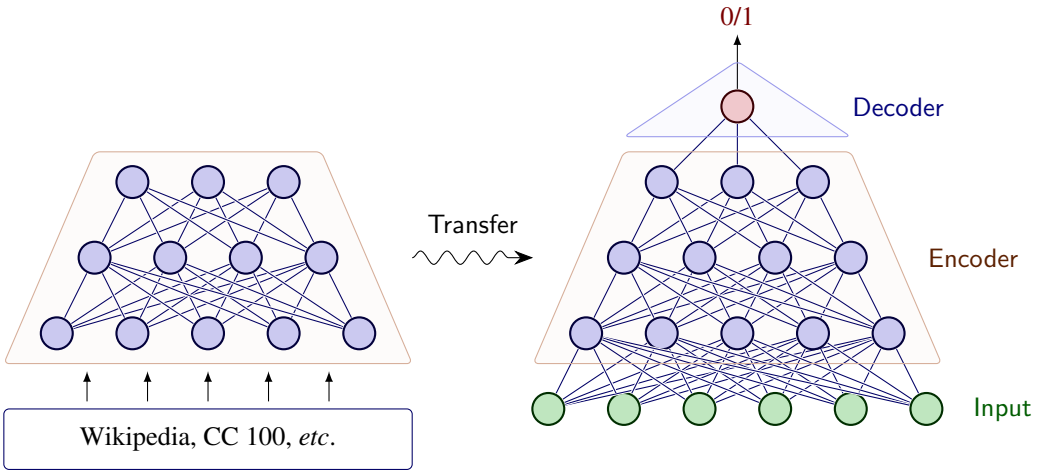

図 2　モデルの概略図．大規模なコーパス（表 1）で事前学習された encoder と，単層の decoder との組み合わせ．入力にはトークン列に変換した説明文を与える．

設定キー	CN	CA	合計
語の最小出現数	68	11	79
上位	54	11	65
描画する共起関係の選択	61	1	62
外部変数	8	34	42
品詞による語の取捨選択	12	1	13
最小文書数	7	1	8
集計単位	7	1	8

表 2　共通の設定キーと出現頻度．CN は共起ネットワークを，CA は対応分析を表す．

設定キーの選択　3 節の分類器 Found_k は本来は分析手法にも依存する概念であるが，データセットを分析手法ごとに分割するとサイズが小さくなり，学習が十分でなくなる恐れがある．したがってデータセット内の共起ネットワーク，対応分析いずれの分析手法にもに対応できるよう，共通の設定キー

$$\mathcal{K}_{\text{共起ネットワーク}} \cap \mathcal{K}_{\text{対応分析}}$$

に対してのみ分類器を学習した（表 2）．

これは「分析手法によらず，同じ設定キーなら似たような説明文（したがって似た分類器）となる」という仮定に基づく選択である．

4.3　結果

交差検証の各 fold においてメトリクス acc を計算し，モデルごとに集計した．メトリクスの分布の五数要約を図 3 に示す．

五数要約のいずれの指標においても，BERT によるモデルが最も高い精度を達成できていることが確かめられる．特に最小値および第 1 四分位数において大きな精度差が見られる．このことは，我々の提案モデルが一定以上の精度を維持することができ，その意味で安定していることを意味する．

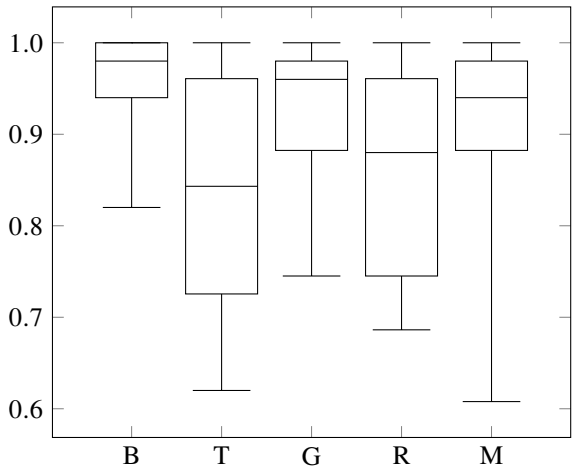

図 3　評価結果．acc の分布についての五数要約を箱ひげ図で並べた．横軸のラベルはそれぞれ BERT，T5，GPT-2，RoBERTa，森田らによるモデルを表す．

5　おわりに

テキストアナリティクスツールのログ生成タスクについて，まず分析手法分類，設定キー認識，設定値抽出の三つのサブタスクへの分解ができることを述べた．そのうちの設定キー認識のためのモデルとして，事前学習された言語モデルに基づく設定キーごとの分類器を提案した．その結果，我々の提案モデルは安定して高い精度を記録することができた．

これを設定キーに依存しない一般的な分類器へと汎用化することは，今後の課題となっている．また，ここで触れなかった他のサブタスクに関しても，さらなる研究が必要である．

参考文献

[1] David M. Berry. **Introduction: Understanding the Digital Humanities**, pp. 1–20. Palgrave Macmillan UK, London, 2012.

[2] 森田康介, 西村太一, 亀甲博貴, 森信介. テキストアナリティクスツールの操作ログからの実験設定の説明文生成. In **Proceedings of the Twenty-ninth Annual Meeting of the Association for Natural Language Processing**, pp. 2785–2790, Okinawa, Japan, 2023. 言語処理学会.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **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.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[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, 2019.

[7] 樋口耕一. テキスト型データの計量的分析. 理論と方法, Vol. 19, No. 1, pp. 101–115, 2004.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[9] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[10] Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1703–1714, Online, July 2020. Association for Computational Linguistics.

[11] Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim, 2019. Leibniz-Institut f'ur Deutsche Sprache.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.