ラベル依存の注視機構を用いた医薬系文献に対する統制索引語付与

菊井 玄一郎　　鈴木 慶二　　関根 基樹　　水田 寿雄
国立研究開発法人　科学技術振興機構(JST)　情報企画部
{genichiro.kikui, k3suzuki, sekine, mizuta}@jst.go.jp

概要

本稿では、深層学習に基づくテキスト分類手法を用いた医薬系文献に対する統制語索引付与について報告する。本稿で試みた手法は基本的には LSTM および BERT によるマルチラベル・テキスト分類であるが、既存研究に従い、「ラベル依存の注視機構」を利用している。専門分野の文献を扱うため、当該分野のコーパスによる BERT 事前学習や単語埋め込みの構築も試みた。F1 値により性能評価を行ったところ、ラベル依存の注視機構を備えた LSTM が同様の BERT より若干上回る結果となった。またテキスト表層から付与しにくい索引も一定程度付与できることが分かった。

1. はじめに

文献に対してその内容（主題）を表す索引語（index term）の集合を付与することは文献検索やテキストマイニングなどにおいて極めて重要であり、特に専門用語シソーラスによる索引は専門分野の文献調査において有用であることが指摘されている[1]。

科学技術振興機構（JST）では文献 DB（メタデータ）整備の中心業務として、約 37,000 語のシソーラス、115 万語の「大規模辞書（用語辞書）」に基づいた索引を各文献に付与している。付与作業は JST の前身組織の時代から専門作業者が行ってきたが、近年、一部文献についてルールベースの自動処理を適用している。しかしながら、管理コストやの点から医薬系分野を含む全面的な採用には至っていない。

そこで、本稿では医薬系分野の文献に統制語索引を付与する手法を検討する。上述のように JST では相当量の文献に人手による索引が既に付与されており、その一部が利用可能であることから機械学習的なアプローチを試みる。

以下では、まず２章で JST の索引について簡単に紹介し、3 章で関連研究について述べる。4 章で我々が試みた方法について説明したあと、5 章で実験データと実験設定について述べ、6 章で結果とその考察を述べ、最後にまとめる。

2. JST の文献 DB における索引

JST の文献 DB において各文献に付与されている索引の種類を図 1 に示す。索引は、当該文献の主題を表すメインヘディング、メインヘディングの補足として医薬系分野で付与されているサブヘディングからなる。本稿ではこれらのうちメインヘディングのみを対象とする。

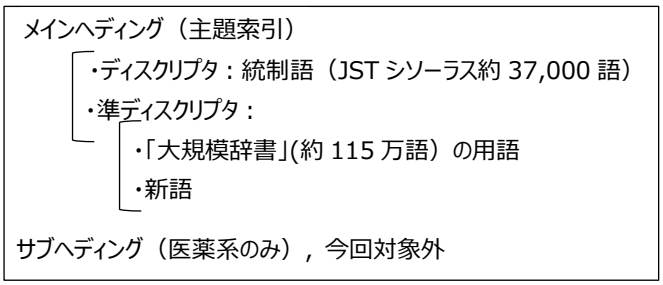

図 1: JST 文献 DB における索引

メインヘディングはさらに現在 37,444 語の「JST シソーラス」を統制語とする「ディスクリプタ」と、それ以外の「準ディスクリプタ」に分けられる。後者についても新語などを除いて、基本的には約 115 万語（同義語 ID 数約 22 万）の「大規模辞書」から選ぶことが推奨されている。「大規模辞書」を統制語と呼ぶのは不適当かも知れないが、極力統制された用語で索引付けを行っているといえる。索引や辞書に関する詳細は文献[2],[3]などを参照されたい。

3. 関連研究

索引の自動付与手法は「抽出型」「分類型」「生成型」の大きく３つに分類できる。

抽出型はテキスト内で「主題表現」や「キーフレーズ」に相当する文字列を特定し、必要に応じて文字列の標準化や統制語へのマッピングを行うというものであり、1950 年代からの長い歴史がある。

分類型は統制語をテキストの「主題」を示す「分類ラベル（カテゴリ）」とみなして、テキスト自動分類の手法を適用するというものである。

最後の生成型はアブストラクト型の文書要約と同様に索引語の並びを生成するもので、出力側の語彙を限定すれば統制型の索引も可能である。

これらのうち本稿ではテキスト全体から索引の集合を直接推定する「分類型」に焦点を当てる。JST 文献 DB における主題索引の約 4 割が同義語を含めてもテキスト中に出現しないことが指摘されており[1]、分類型が一つの解決策と考えたからである。

分類型の索引付与はテキスト分類と基本的に同じタスクとみなせるが、１）一つの文書に対して複数の分類ラベル（索引）を選ぶ「マルチラベル分類」であり、２）ラベルの数がかなり多い（数万～数百万）ことが特徴である。このようなタスクは大規模マルチラベル分類(Large-scale Multi-label Text Classification: LMTC)と呼ばれ、例として、診療記録に診断・治療を表す ICD-9[i]の 8,771 ラベルを割り当てる MIMIC-III[4]、法律文書に法的な概念(4,271 ラベル)を割り当てる EUR-Lex[5]、商品説明文に約 13,000 ラベルの商品カテゴリを割り当てる Amazon13K[6] などがある。さらにラベル数が多いものは「超大規模マルチラベル分類(eXtreme Multi-label text Classification: XMC)」と呼ばれ、Amazon の 67 万ラベルのデータセットなどがある[ii]。今回、我々が取り組む課題は分野内容的に上記 MIMIC-III に近くラベル数は 12,000 程度である。

LMTC, XMC については深層学習を利用した多くの研究が行われている。転移学習型手法の普及以前は CNN や RNN に「ラベル依存の注視機構（Label-Wise Attention Networks: LWAN）」を組み合わせた方法が試みられた[7],[5]。また、大量のラベルに起因するデータ過疎性の問題に対応するために、ラベルの階層構造を利用する方法[8]や、自動クラスタリングによりラベルを階層化し、少数カテゴリに対するマルチラベル分類をトップダウンに適用する方法などが提案されている[9][10]。特に AttentionXML[10]は分類処理として LSTM(Long-Short Memory)にラベル依存の注視機構を組み合わせた手法(LSTM-LWAN)を用いることにより、精度向上を実現している。BERT[11]の提案を受けて、文献[12]では BERT にラベル依存の注視機構を組み合わせた BERT-LWAN と呼ぶ手法を提案し、上述の 3 つの LMTC タスクで既存手法との比較実験を行った。各タスクで安定して性能が高かったのは AttentionXML[10]であり、MIMIC-III で BERT-LWAN に 19 ポイントの差をつけて 1 位、EURLEX, AMAZON13k でも 1 位の BERT-LWA より 2-3 ポイント低い程度であった[iii]。なお、MIMIC-III において BERT-LWAN の性能が低い理由として、事前学習モデルが汎用であったこと、入力テキスト長がモデルのサイズ(512tokens)を超えていたこと、サブワードによる専門用語の過度な分割などを指摘している。

以上より、AttentionXML(LSTM-LWAN)や BERT-LWAN などのラベル依存の注視機構が有効であること、BERT を利用する場合は適用先ドメインのコーパスによる事前学習が有効であると考えられる。

4. 試みたマルチラベル分類手法

本節では本研究で適用を試みた２つの手法（モデル）について説明する。

4.1 ラベル依存注視機構付き LSTM

ラベル依存の注視機構付き LSTM(LSTM-LWAN)（図２）は AttentionXML[10]の分類処理で利用されている手法である。双方向 LSTM の各トークンに対する出力（埋め込みベクトル）にラベル依存のアテンション重みを掛けて総和したものを当該ラベルに対する文書全体の pooling ベクトルとする。具体的には、j番目のラベルに対するi番目のトークンの重みα_{ij}(スカラー値)を次式で計算する。

$$\alpha_{ij} = \frac{\exp(\boldsymbol{h_i} \cdot \boldsymbol{w_j})}{\sum_{t=1}^{T} \exp(\boldsymbol{h_t} \cdot \boldsymbol{w_j})}$$

[i] World Health Organization’s Ninth Revision, International Classification of Diseases (ICD-9)

[ii] http://manikvarma.org/downloads/XC/XMLRepository.html

[iii] 文献[10]の実験ではラベル数 4 万以下の LMTC では階層化を行っていない。文献[11]の実験設定は不明である。

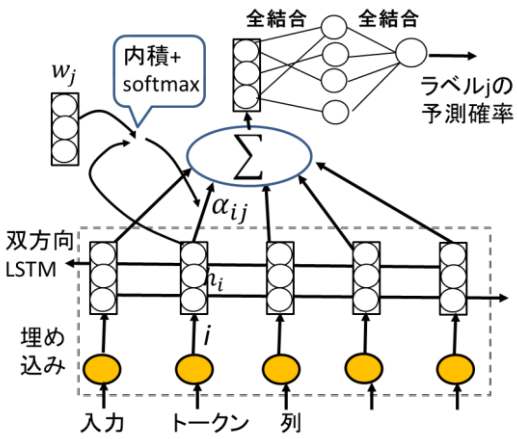

図 2: LSTM-LWAN の概要
(BERT-LWAN は点線枠を BERT に置き換えたもの)

ここで h_iは i 番目のトークンに対する LSTM の出力（埋め込み）、w_jはj番目のラベルに対する注視係数ベクトルであり、学習で決まる。j番目のラベルに対するテキスト全体の pooling ベクトルm_j は次式の通り各トークンの埋め込みの加重和である。

$$m_j = \sum_{i=1}^{T} \alpha_{ij} h_i$$

このベクトルを通常のマルチラベル分類と同様、全結合層（中間層を 1 層入れている）を通してラベルごとに 0/1 の出力を行う。なお全結合層の重みは全てのラベルで共通としている。

4.2 ラベル依存の注視機構付き BERT

4.1 節の LSTM と埋め込み（図 2 の点線枠）を BERT に置き換えたものが BERT-LWAN[4]である。すなわち 4.1 節のh_iの部分がi 番目のトークンに対する BERT の埋め込み出力となる。

5. 実験

5.1 実験データ

5.1.1 分類実験(finetuning)用データ

分類実験用データとして JST が整備している文献 DB のうち医薬系文献の一部を用いた。対象とする索引語はカテゴリーコード LS44(薬物)、または、LS51(病気・病理・症状)に分類される主題索引（メインヘディング）であり、大規模辞書を用いて動機語を代表表記に集約した。利用した文献データは医薬系文献 DB である JMEDPlus に収録されている文献のうち上述の索引語を含み 2018 年から 2021 年 8 月までに作成された約 38 万文献であり、8:1:1 の比率で訓練、検証、評価用にランダムに分けた。このうち 18 万文献（全て口頭発表の予稿）については本文データが存在するのでそれらも入力テキストとして利用した。文献数とラベル数(type 数)を表 1 に示す。索引（語）を 2 つのカテゴリに絞っているため、1 文献あたりの正解ラベル数（索引数）の平均は 5.5 個である。正解ラベルのうち JST 大規模辞書による同義語展開を行っても抄録・本文・表題の文字列に含まれないものを「本文未出現ラベル」として同定した。本文未出現ラベルは 1 文献あたり平均 2.1 個(39%)であった（ラベル数 0 の文献も分母に含む）。

表 1　索引実験用のデータ

	訓練	検証	評価
文献数	309,878	34,431	38,257
ラベル種類	13,230	9,803	10,022

分類処理への入力テキストは上記文献 DB 中の表題と（あれば）抄録、本文が存在するものは本文をこの順にスペースでつないだものである。入力テキストの文献あたりの平均文字数は 409 であった。

5.1.2 事前学習用データ

事前学習用のコーパスは医薬系文献を含む JST 科学技術文献 DB の抄録 960 万文献(1 抄録あたり平均 4.2 文、240 文字)を用いた。なお、分類の評価用、検証用の文献は含んでいない。

5.2 実験設定

5.2.1 分類モデルの構築

分類モデルは 4 節で述べた通りである。損失関数は通常のマルチラベル分類と同様に Binary Cross-Entropy Loss を用いた。ハイパーパラメータは検証用セットで評価値が最高になるものを選んだ。その際の評価尺度は文献ごとの F1 値のマクロ平均である。この種の問題ではラベルごとに F1 値を算出しそれを平均することが多いが、ユーザとしては、文献ごとにどの程度正確に索引されているかが重要であることからこのようにした。

ラベル依存の注視機構の効果を見るために 4 節で述べた 2 つの手法の他にベースラインとして BERT の CLS トークンに対応する埋め込みベクトルを全結

合で出力層に接続したシンプルなマルチラベル分類[iv]も適用した。

5.2.2 LSTM-LWAN 用の単語埋め込み

LSTM の入力層で必要となる埋め込みベクトルは分類学習用のコーパス(5.1.1 節)に GLoVe[13]を適用して 300 次元のものを作成した。形態素解析には MeCab/ipadic に JST シソーラスから作られたユーザ辞書[14]を追加したものを用いた。サブワード分割は行っていない。

5.2.3 BERT 事前学習モデルの構築

BERT の事前学習モデルは 5.1.2 で述べたコーパスから作成した。まず、5.2.2 で述べた JST シソーラス由来のユーザ辞書を用いて形態素解析し、これをもとに語彙サイズ 32K のサブワードモデルを作成した。このサブワードモデルと既存ツール[v]を用いて事前学習を行った。モデルのハイパーパラメータは東北大 base モデルに合わせて隠れ層 12 層 768 次元、ヘッド数 12 などとした。NVIDIA RTX A6000x8 台で 14 日間訓練した。

6. 実験結果と考察

評価セットに対する分類性能（索引性能）を表 2 に示す。事前学習モデルの base は東北大 base model, JST は 5.2.3 で述べたモデルである。prec(精度)、rec.(再現率)、F1 値は、前述の通りそれぞれを各文献について求めた上で全文献の平均を取った値（マクロ平均値）である。なお、分母が 0 になる場合は評価値を 0 として計算した。rec2 は正解ラベルを「本文未出現ラベル」に限定した場合の再現率であり（本文未出現ラベルを含む文献 28,932 文献で計算）、F1 値上位２つの設定で計算した。目的は表層一致を超えた索引抽出能力を評価するためである。

表 2　分類（索引）性能

分類モデル	事前学習モデル	評価セットスコア			
		prec.	rec.	F1	rec2.
BERT_MLC	base	**.726**	.517	.561	-
	JST	.716	.548	.581	-
BERT_LWAN	base	.615	.624	.575	-
	JST	.658	.660	.617	**.493**
LSTM-LWAN	-	.675	.675	**.632**	.473

まず、F1 値については AttentionXML が２位の JST 事前モデルを用いた BERT-LWAN より 2 ポイント高い結果となった。この順位関係は先行研究[12]と一致する。BERT-LWAN について、今回は索引対象と同じ分野のコーパスで事前学習モデルを訓練していることから事前学習モデルのミスマッチは少ないと考えられる。他の理由としては BERT がサブワード単位であるのに対して LSTM-LWAN が JST 辞書の語彙（専門用語）を GLoVe で直接モデル化していることなどが考えられるが、更なる検討が必要である。

抽出型の索引付与手法が基本的にはテキストに出現する表現の中から索引表現を探す処理であるのに対して、今回試みた分類型は対象テキストの表層にない索引語を付与できる可能性がある。rec2 を見ると、再現率 0.5 弱になっており、全ての正解索引に対する再現率 0.66 より 18 ポイント程度下がるものの全体の適合率(0.66)を考えるとある程度できていることが分かる。なお、BERT-LWAN の方が LSTM-LWAN より若干良いことを注記する。

BERT については科学技術分野のコーパスで作成した事前学習モデルは汎用のものと比べて性能が高く、BERT-LWAN で 4 ポイント程度向上している。また注視機構は BERT に対して 2-4 ポイントと性能向上に一定の寄与があったと考えられる。

7 おわりに

本稿では統制語索引を文書分類的に解く方法について、JST の文献 DB におけるメインヘディング付与を題材として、特にラベル依存の注視機構(label-wise attention:LWAN)の適用を試みた。

実験結果によればラベル依存の注視機構は有効であり、F1 値として 0.63 程度となった。LWAN を、対象分野のコーパスから作成した GLoVe、および、LSTM と組み合わせる方法が BERT と組み合わせた方法より若干上回る性能となった。

索引付けについてはテキスト中から重要表現を取り出す抽出型が基本であり、今回試みた分類型の方法はそれを補足する役割があると考えている。今後は両手法の改善に加え、それぞれの性質の分析を踏まえて適切に組み合わせることが大きな方向性であると考えている。

[iv] Huggingface の BertForSequenceClassification をマルチラベル分類の設定で使用。

[v] Huggingface の BertForMaskedLM

参考文献

[1] 堀内美穂, 中村徹, 永井賢吉: JSTDB 検索における索引の有効性と索引作業の重要性-JSTPlus ファイル, JMEDPlus ファイルにおける索引語の分析-, 情報管理, vol.52, No.1, 2009.
[2] 冨永祥平: “索引作業におけるより高度な支援辞書の利用～JST 抄録・索引支援システム「NAISS」について～”, 情報管理, vol. 50, no. 4, pp. 210-217, 2007.
[3]山崎文枝, “JST 科学技術シソーラスの改訂”, 情報管理, vol.60, no.5, pp.365-368, 2017.
[4] Johnson, A EW., et al., “MIMIC-III, a freely accessible critical care database”, Nature, 2017.
[5] Chalkidis, I., et al., “Large-Scale Multi-Label Text Classification on EU Legislation”, ACL 2019, 2019.
[6] McAuley, J., et al., “Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text”, ACM Conf. on Recommender Systems, pp. 165-172, 2013.
[7] Mullenbach, J., et al., “Explainable Prediction of Medical Codes from Clinical Text”, NAACL2018, 2018.
[8] Rios, A., et al., “Few-Shot and Zero-Shot Multi-Label Learning for Structured LabelSpaces, EMNLP2018, 2018.
[9] Khandagale, S., et al., “Bonsai – Diverse and Shallow Trees for Extreme Multi-label Classification”, arXiv:1904.02349v2, 2019.
[10] You, R., et al., “AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification”, NeurIPS2019, 2019.
[11] Devlin, J., et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proc. NAACL2019, 2019.
[12] Chalkidis, I., et al., “An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels”, EMNLP2020, 2020.
[13] Pennington, J. et al., “Glove: Global vectors for word representation”, EMNLP2014, pp1532-1543, 2014.
[14] 建石由佳, 信定知江, 高木利久: “JST 科学技術用語シソーラスに基づく MeCab 用専門用語辞書”, 言語処理学会第 23 回年次大会、P7-1 (pp485-488）, 2017.