sequence-to-sequence モデルを用いた疾病表現への大規模医療コードの自動付与

畑中 裕貴　　秋葉 友良
豊橋技術科学大学
hatanaka@nlp.cs.tut.ac.jp, akiba@cs.tut.ac.jp

1 はじめに

ICD-10(The International Classification of Diseases, tenth Edition) は疾病や医療処置を扱った分類システムの 1 つで、約 70,000 件のコードで構成される。自然言語処理の分野ではこれまでに、医療テキストデータへの ICD コードの付与に関する研究がいくつか行なわれてきたが、学習データのスパース性のために、学習データに含まれないような珍しい疾病を表すコードを扱うことが困難であった。それゆえに先行研究 [1] [2] では、扱うコードの数を学習データ中の出現頻度などで制限している。

本論文では、sequence-to-sequence (seq2seq) モデルを用いた生成による ICD-10 コード付与手法を提案する。seq2seq モデルは似た疾病を表す ICD-10 コード間の関係を利用することで、ICD-10 コードの文字シーケンスに見られる規則性を自動的に学習する。このことにより、提案手法は学習データに含まれないコードを生成し付与することが可能である。実験では、ICD-10 コードと対応するディスクリプションのペアをランダムに 2 つのグループに均等に分割し、1 つのグループを用いて、ディスクリプションのトークン系列から ICD-10 コードを表す文字系列に変換する seq2seq モデルの学習し、もう 1 つのグループを用いてコード予測テストを行なった。同じデータで学習した従来の分類器モデルと提案手法の比較を行なった。

加えて、医療現場でのコード付与を想定し、自然発話データからのコード予測テストを行った。最後に、seq2seq モデルが存在しないコードを生成してしまうことを防ぐ手法を提案する。

2 関連研究

ICD コードの付与に関する研究では近年、ディープニューラルネットワークを用いた手法が盛んになっている。Tai ら [1] は 6,500 件の ICD-9 コードと 1,047 の 3 桁目までの ICD-9 コードの付与実験を HA-GRU モデルを用いて行なった。Xie ら [2] は 2,833 件の ICD-9 コードの付与実験を Tree-LSTM 用いて ICD コード間の階層構造を明示的にモデルに与えて行なった。これらの研究では、ICD-9 という ICD-10 よりもコード数が少ないものを使用したり、扱うコードの数を制限したりすることで、珍しい疾病に関する学習データが不足しているスパース性の問題に対処している。

一方、本研究では約 70,000 件からなる ICD-10 コード全てを対象とする。提案手法は seq2seq モデルを用いることで、学習データに含まれていないコードを扱えるようにした。

3 手法

3.1 ICD-10

ICD-10 は約 70,000 件のコードで構成されており、体系的にまとまられている。各コードの文字系列には ICD-10 の持つ階層構造が反映されている。コードの上位の桁は粗い分類を表しており、反対にコードの下位の桁はより細かい分類を表している。例えば、”K” で始まるコードは ”digestive system（消化器系）” の病気を表している。” K” の次に ”4” が続くコードはヘルニアを表す。ヘルニアは ”digestive system ” の病気の一つである。”K4 ” で始まるコードは ” K ” で始まるコードの部分集合となっている。ICD-10 コード ” K41.41 ” は Unilateral femoral hernia, with gangrene, recurrent ” を表している。

3.2 seq2seq モデル

本論文では、seq2seq モデル [3] を用いた生成による ICD-10 コード付与手法を提案する。エンコーダとデコーダには GRU [4] を使用した。図 1 にモデルの構成を示す。入力と出力はそれぞれ、コードのディスクリプション（のトークンの系列）とコードの文字系列である。提案手法ではコードの文字系列が持つ階層構造に従って、各タイムステップでデコーダはその前の（より粗い）予測結果に基づいて 1 文字を予測する。

従来の分類器モデルは最も細かい分類を 1 度に予測するため対照的である。提案手法は似た疾病を表すコード間の関係を利用し、ICD-10 コードの文字シーケンスに見られる規則性を自動的に学習することで、学習データに含まれていないコードをも扱うことができる。例えば、コードのディスクリプションの末尾に ”, unspecified ” が含まれる場合、その ICD-10 コードの最後の桁は ” 9 ” であるという規則性がある。

また、1 度提案手法のモデルを学習したら、それを様々なレベルの分類予測に応用できる。例えば seq2seq モデルの生成したコードに関して、最初の 3 桁が一致したものを 1 つのクラスとみなせば、3,120 クラス分類の結果を得ることができる（1〜 3 桁目はそれぞれ 26, 10, 12 種類の文字パターンが考えられる）。

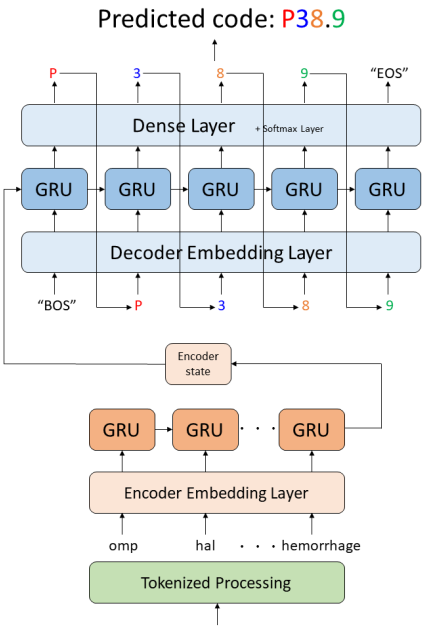

図 1: 提案モデルの構成

ここで入力のトークンの系列を表現するベクトル X を $[x_1,x_2,...,x_i,...,x_I]$、出力の文字の系列を表すベクトル Y を $[y_1,y_2,...,y_j,...,y_J]$ とする。x_i と y_j は one-hot ベクトルで、その次元サイズは $|V_X|$ と $|V_Y|$ となる。ここで V_X と V_Y はそれぞれ入力と出力の語彙である。

エンコーダにおいて、入力系列の隠れ表現は、トークンの並びを考慮しながら生成される。ここで、E^{enc} はエンベディング行列で、最終隠れ状態ベクトル $State_I^{enc}$ はデコーダに送られる。

$$Emb_i^{enc} = E^{enc} x_i$$
$$Output_i^{enc}, State_i^{enc} = GRU(Emb_i^{enc}, State_{i-1}^{enc})$$
$$State_0^{dec} = State_I^{enc}$$

デコーダにおいて、GRU は $Output_j^{dec}$ を 1 つ前のタイムステップ j-1 の情報を使いながら出力する。ここで、E^{dec} はエンベディング行列で、$Input_0$ は文頭 (BOS) を表す one-hot ベクトルである。

$$Emb_j^{dec} = E^{dec} Input_j^{dec}$$
$$Output_j^{dec}, State_j^{dec} = GRU(Emb_j^{dec}, State_{j-1}^{dec})$$
$$Pre_j = Softmax(Dense(Output_j^{dec}))$$

3.3 attention モデル

attention は seq2seq モデルに導入され、機械翻訳などの分野で性能の改善を示している [5]。デコーダを次のように修正した。

$$Outputs^{enc} = [output_1^{enc}, output_2^{enc}, ..., output_I^{enc}]$$
$$Emb_j^{dec} = E^{dec} Input_j^{dec}$$
$$Att_j = Attention(Outputs^{enc}, Outputj^{dec})$$
$$Attention = tanh(W^{(a)}[Output_j^{dec}, \bar{h}])$$
$$\bar{h} = \sum_{i=1}^{I} a_i Output_i^{enc}$$
$$a_i = Softmax(Output_i^{enc} \cdot WOutput_j^{dec})$$
$$Pre_j = Softmax(Dense([Att_j, Outputj^{dec}]))$$

ここで、$W^{(a)}$ と W は学習可能な行列で、a_i は入力系列の i 番目のトークンと出力系列の j 番目の文字との類似度になっている。

3.4 生成確率の利用

提案手法は seq2seq モデルのデコーダの各タイムステップで 1 文字ずつ予測を行なうため、最終的に存在しないコードを生成してしまうことがある。この問題に対処するために、存在するコードの seq2seq モデルにおける生成確率を利用する手法を提案する。存在するコード全ての seq2seq モデルにおける生成確率を計算し、最も生成確率が高いコードをこの手法の出力とする。

入力系列 X が与えられたとき、あるコードを表す系列 Y が生成される確率 $P(Y|X)$ は次の式のように表される。

$$P(Y|X) = \prod_{j=1}^{J} P(y_j|Y_{<j}, X)$$

4 実験

以下の 2 つの実験を行なった。

- **実験 a:** ICD-10 コードのディスクリプションのトークン系列から、対応するコードの予測およびクラス分類を実験した。
- **実験 b:** 実際の医療現場でのコード付与を想定し、医療データから得られた自然発話に近い表現に対してコードを予測する実験を行なった。

4.1 データ

69,823 件の ICD-10 コードを実験に使用した。各コードには自然言語で記述された 1 つのディスクリプションが存在する。簡略化のため、コードから文字”.”を取り除いた。これらは実験 a では学習とテストに用い、実験 b では学習のみに用いた。実験 b でテストに用いるデータは [6] における自然言語で記述された日本語の医療記録に基づいており、プロの翻訳サービスによって自然発話風に英訳された [7]。疾病表現にアノテーションされている箇所とアノテーション箇所に割り当てられている ICD-10 コードのペアを利用した。ペアの重複や曖昧なコードを含むものを除去し、91 件のデータを用意した。

4.2 実験設定

実験 a では、ICD-10 コードと対応するディスクリプションのペアをランダムに 2 つのグループに均等に分割し、1 つのグループを用いて、ディスクリプションのトークン系列から対応する ICD-10 コードを生成または分類する各モデルの学習を行なった。もう 1 つのグループを用い、学習データに含まれない 34,912 件のコードの予測とコードの最初の 3 桁が一致したものを 1 つのクラスとみなした 3,120 クラス分類の予測の精度をテストした。まず seq2seq モデル、seq2seq モデル+attention（attention モデル）、分類器モデルの性能比較を行なった (実験 a-1)。seq2seq モデルの生成確率を利用する手法に関して、予測時における時間的コストの問題から、上記のデータを用いてテストすることができなかった。そのため、全ペアからランダムに 1,000 件抽出し、残りのペアを用いて seq2seq モデルを学習し、1,000 件のデータでコードの予測精度をテストした。通常の手法と、予測時に生成確率を用いる手法の比較を行なった（実験 a-2）。実験 b では全ペアを用いて seq2seq モデルと分類器モデルの学習を行ない、コードの予測精度に関して seq2seq モデル（2 手法）と分類器モデルの性能比較を行なった。

入力語彙 V_X は BPE によってトークン化されている。BPE による語彙サイズ $|V_X|$ は 5,000 に実験的に決定された。出力語彙 V_Y は文頭と文末を表す特殊記号と ICD-10 コードを構成する数字とアルファベットで構成される。各モデルの学習で、バッチサイズは 300、epoch 数は 300 に設定し、Adam optimizer を使用した。

seq2seq モデルでは Emb^{enc} のエンベディングベクトルの次元数は 500、 Emb^{dec} では 50、GRU の隠れ状態ベクトルの次元数は 300 で、行列 $W^{(a)}$ と W のサイズは 300*300 と 300*600 とした。ベースラインとなる分類器モデルは seq2seq モデルと同じエンコーダと、全結合層、softmax 層で構成される。分類器モデルの出力次元数は分類クラスに依存する。従って分類器モデルにとってコード予測は ICD-10 コードの総数と同じ 69,823 クラス分類となる。実験 a ではトレーニングセットの 10%がバリデーションに使用され、モデルが学習セットに含まれないコードを予測するのに十分な学習が行なわれていることを確認した。

評価指標には Accuracy(%) を用いた。これはテストサンプルのうち、どれだけが正確に予測されたかを表す。

4.3 実験結果

実験 a の結果を表 1 と表 2 に示す。表 1 より提案手法は、全 ICD-10 コードの半分のデータだけで学習を行っても、学習データに含まれていないコードの 80%を正確に予測できた。これは seq2seq モデルが ICD-10 コードの階層構造や割り当てルールを学習し、そのルールに従ってコードを生成したからであると考えられる。この結果は提案手法が、あるクラスに関する情報が学習データで不足していたとしても、その階層構造や学習データのもつルールを用いることで一定の精度で予測することが可能であることを示す。分類器モデルでは 1 件も予測することができなかった。3,120 クラス分類では、seq2seq モデルは結果こそ分類器モデルに劣ったが、この分類問題を解くためのモデルでないにもかかわらず、分類器モデルと同等の結果を示した。重要なのは、分類器モデルは 3,120 クラス分類や 69,823 クラス分類をそれぞれ学習しなければならないが、seq2seq を用いる提案手法は、さまざな粒度の分類に対して 1 回の学習で済むという点である。表 2 より、生成確率を使用する手法では、存在しないコードを生成することがなくなり、Accuracy の値が改善した。

実験 b の結果を表 3 に示す。提案手法が Accuracy の値で分類器モデルを上回るという結果が得られたが、実験 a などに比べ、Accuarcy の値が大きく減少

表 1: 実験 a-1 の結果

	Acc(%)	
	コード予測	3,120 クラス分類
seq2seq モデル	**79.7**	89.3
attention モデル	78.3	88.4
分類器モデル	0.0	**90.9**

表 2: 実験 a-2 の結果

	Acc(%)
	コード予測
seq2seq モデル	85.9
seq2seq モデル（生成確率を使用)	**86.8**

してしまった。これは学習データの入力であるコードディスクリプションとテストデータの入力の書式に差があるためであると考えられる。GAN [8] の導入などにより、書式の差に対応する仕組みを検討したい。また、提案手法は入力“small feces”から存在しないコード”R155”を生成したが、生成確率を使用する手法では正しいコード”R195”を予測することができ、Accuracy の値が改善した。

表 3: 実験 b の結果

	Acc(%)
seq2seq モデル	13.2
seq2seq モデル (生成確率を使用)	**14.2**
分類器モデル	12.1

5 おわりに

本論文では seq2seq モデルを用いた生成による ICD-10 コード付与手法を提案した。実験では、提案手法は学習データに含まれないコードでさえ予測することができた。これは提案手法がデータのスパース性によって問題が発生する大規模なクラス分類において有効であることを示した。従来の分類器モデルは学習データに含まれないコードを予測することはできなかった。さらに、seq2seq モデルから生成されたコードを直接用いて、分類問題を調査し、分類器モデルとの比較を行なった。提案手法は 1 回学習しただけで、様々な粒度の分類問題に適応できる。

医療データから得られた自然発話に近い表現に対するコード予測では、コードのディスクリプションを入力とした場合に比べて予測精度が大きく減少した。今後の課題として、実際の医療現場でのコード付与に応用できるよう、ICD-10 のディスクリプションのような公式な表現と自然発話に含まれるような疾病表現との差の解消を行なう学習アーキテクチャの設計に取り組みたい。

謝辞

本研究は JSPS 科研費 19K11980 の助成を受けた。

参考文献

[1] Tal Baumel, J. Nassour-Kassis, R. Cohen and M. Elhadad. Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment. In Proc. of AAAI, 2018.

[2] P. Xie, E. Xing and M. Young. A Neural Architecture for Automated ICD Coding. In Proc. of ACL, 2018.

[3] I. Sutskever, O. Vinyals, and Q.V. Le. Sequence to sequence learning with neural network. In Proc. of NIPS, 2014.

[4] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proc. of EMNLP, 2014.

[5] D. Bhdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, In Proc. of ICLR, 2015.

[6] E. Aramaki, M. Morita, Y. Kano, T. Ohkuma, Overview of the NTCIR-11 MedNLP-2 Task, In Proc. of the 11th NTCIR Conference, 2014.

[7] T. Akiba, B. Sy and A. Zeidan. ICD-10 code retrieval based on distributional semantics of diagnosis descriptions. In Proc. of ICAICTA, 2017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair et al. Generative adversarial nets. In Advances in neural information processing systems. In Proc. of the 27th International Conference on NIPS Vol. 2, 2014.