5W-05

Lambda-network を用いたエッジデバイスのための低遅延な対話状態追跡

○石島 侑弥 † 李 晃伸 † 西山 達也 † 古賀 光 ‡ 西島 敏文 ‡ 佐々木 悟 ‡

名古屋工業大学 † トヨタ自動車 ‡

1 はじめに

近年の NN モデルは大規模なモデルが多く, NN ベースの対話システムも大規模化が進んでいる．そのため，システムの多くがクラウド上で動作される．また，近年エッジコンピューティングと呼ばれる技術が注目されている．エッジコンピューティングは，携帯端末や車載端末などユーザに近い端末 (エッジデバイス) でデータ処理を行うことで，通信遅延の改善，データ漏洩のリスク低減に貢献できる．つまり，対話システムをエッジデバイス上で動かせば，現行のシステムより高いセキュリティで，早い応答が可能になる．従って本研究では，エッジデバイス上での対話システムの動作を目指し，対話システム中の対話状態追跡モデルを軽量化する．

対話状態追跡とは，発話中から意図や要求を抜き出して対話状態を更新・保持する技術である．例えば,「中華を食べたい」という発話ならば,「意図：検索, 種類：中華」と対話状態を更新する．本研究で使用する BERT[1] を用いたベースラインモデルは，対話履歴から直接対話状態を出力するため, 対話の流れは対話履歴から捉える．つまり，対話履歴を多く入力するモデルは，より対話の流れを捉えた高性能なモデルとなる．しかし，対話履歴を多く入力すれば，モデルの最大入力長が増え，計算量が増加するため，エッジデバイス上での高速・低遅延な動作が難しくなる．今回，履歴

Low-latency dialogue state tracking for edge devices using Lambda-network

† Yuya Ishijima, Akinobu Lee, Tatsuya Nishiyama, Nagoya Institute of Technology

‡ Ko Koga, Toshifumi Nishijima, Satoru Sasaki, Toyota Motor Corporation

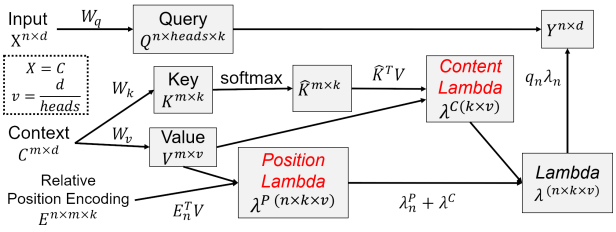

図 1 Lambda-network のモデル図

数増加による計算量増加は，最大入力長の 2 乗という計算量がかかる BERT 中の Self-Attention 機構の影響が強いと考えた．そこで本研究はその Self-Attention を，文脈情報を固定サイズの線形関数に要約する Lambda-network[2] に置き換えることで，履歴数増加による計算量増加を抑えつつ，軽量で高精度な推論が可能な対話状態追跡モデルを作成する．

2 提案手法

Self-Attention は，入力文中の各単語の関連度スコアを計算し，単語ベクトルに掛け合わせることで，どの単語同士に依存関係があるのかを捉えられる．そのため，対話文などの長い Context に対しても，長距離依存関係を捉えた埋め込みベクトルを生成することができる．

Lambda-network は，図 1 のように 2 つの Lambda 行列を作成することで，Self-Attention に比べて少ない計算量・メモリ使用量で，長距離依存関係を捉えた埋め込みベクトルを出力できる．1 つ目は Content-Lambda と呼ばれ，$k \times v$ の固定サイズで Context の文脈情報を要約して捉える行列，2 つ目は Position-Lambda と呼ばれる n 個の位置ベクトルで，Context 中の各単語の位置情報を捉える行列である．この 2 つを足した Lambda を Query に掛け合わせることで，Self-

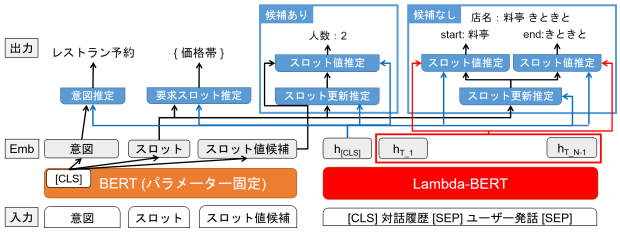

図 2　Lambda-BERT を用いた対話状態追跡モデル

Attention 同様に文脈情報と各単語の依存関係を捉えた埋め込みベクトルを出力できる．また，Lambda の演算に用いられる Query, Key, Value のサイズも k や v に圧縮されるため，Y の出力にかかる計算量・メモリ使用量が削減されている．本研究では，BERT の Self-Attention 部を全てこの Lambda-network に置き換えて，BERT と同じ手法で事前学習を行う．そして，その事前学習モデルを，ベースライン同様に Embedding 部に用いて Fine-Tuning を行う (図 2)．

3　実験

今回は，SGD データセット [3] の 642 対話 (9954 発話) を日本語に翻訳したデータセットを用いて，対話状態追跡の精度を示す Joint Goal Accuracy で評価を行った．比較モデルの Embedding 部には，日本語版 BERT である東北大 BERT[*1] を用いた．また，Lambda-network のハイパーパラメータである k と v は，Lambda-network の論文 [2] と東北大 BERT を参考にして，$k = 16, v = 64$ とした．

まずパラメータ数に関しては，東北大 BERT が 1.1 億なのに対し，Lambda-BERT は 0.85 億と 22.7 % の削減ができている．精度に関しては，表 1 から，2, 3 発話入力ならば精度低下は 0.5 から 2.0 % に抑えられるが，4 発話入力以降は差が開くことが分かる．この結果から，文脈情報を固定サイズで表現する Content-Lambda には限界があると考えられる．また，図 3 では，Lambda-BERT が推論時間を最大で 28.6 % 削減しており，最大入力長増加による推論時間の増加量も 3 分の 2 に抑えることができた．これは，圧縮された行列で演算を行う Lambda-network に置き換えたことで，全体的な計算量が減り，また入力長が影響する演算個所を削減できたからである．

[*1] https://github.com/cl-tohoku/bert-japanese

表 1　Joint Goal Accuracy による精度比較

入力発話数	東北大 BERT	Lambda BERT
2 発話	46.4%	44.4%
3 発話	48.2%	47.7%
4 発話	58.8%	47.4%
5 発話	55.8%	47.3%
6 発話	62.4%	52.2%
7 発話	60.7%	54.6%
8 発話	64.5%	55.0%

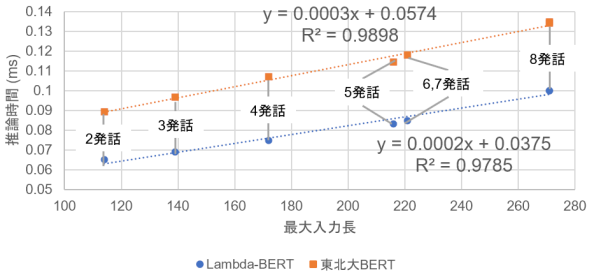

図 3　1 サンプル当たりの推論時間

4　むすび

本研究では BERT の Attention 部を全て Lambda-Network に置き換えた Lambda-BERT を提案した．Lambda-BERT は精度低下を 0.5% に抑え，パラメータ数を 22.7%，推論時間を 28.6 % 削減できるため，計算資源が限られるエッジデバイス上での実行を容易にできる．しかし，入力長が長いと精度の低下が顕著になるため，入力長ごとに，Lambda-BERT のパラメータを調整する必要があると考えられる．

参考文献

[1] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention, 2021.

[3] Abhinav Rastogi et al. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset, 2020.