E-046

質問応答システムにおける各手法の性能比較：BERT が必ずしも有利とならないケースについて
Performance Comparison of Various Methods in Question Answering Systems : About the Case Where BERT is Not Always Advantageous

水口 綾乃† 市川 治†
Ayano Mizuguchi Osamu Ichikawa

1. はじめに

言語モデルにおける課題の一つとして、自然言語での多様な言い回しに柔軟に対応するための表現の確保が存在する。

近年ではこの課題の解決策として、多量の教師なしデータから事前学習を行う手法が主流である。事前学習後モデルの利用方法には、分散表現を取り出して利用する方法や、事前学習したモデル自体を個別タスクへとファインチューニングする方法がある。前者の例としては Word2Vec[1]、後者の例としては最新の言語モデルである BERT[2]がある。BERT では特に文脈を含めた意味の把握が期待できる。

筆者らは、これらを利用しオープンキャンパス用対話システムの性能向上を図った。このシステムはソフトバンク社開発のヒト型ロボットの Pepper を利用して運用しているものである。

システムは、オープンキャンパスに訪れる受験生を利用対象者として想定し、学部や入学試験に関する質問への回答や学内の案内等を行う。

その構成を図 1 に示す。まず Pepper の頭部のマイクロフォンで取得した音声データを Wi-Fi ネットワークで PC へ送り、PC 側で発声区間検知(VAD)を行い音声認識部で音声をテキストに変換する。そのテキストを自然言語理解部で質問クラスに分類する。質問クラスに応じてあらかじめ作成してある応答文を選択し、ロボットの合成音声で返答するという仕組みである。

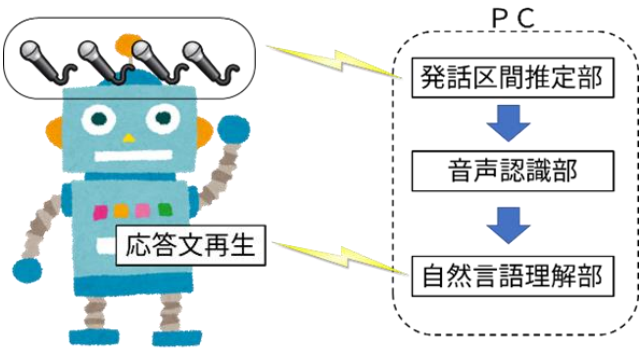

図 1 質問応答システムの構成

この論文では作成したオープンキャンパス用対話システムの自然言語理解部を例に取り、実際の対話システムを構築する際の各種法の性能比較を行う。

† 滋賀大学 Shiga University

2. 関連技術

2.1 BERT

BERT は Devlin らによって提案された、大規模データでの事前学習と小規模な教師ありデータによるファインチューニングにより、自然言語処理タスクを高精度で達成するモデルである。

BERT のモデル構造は Transformer のエンコーダ部分を利用したものであり、Transformer と同じく文脈に合わせた単語のベクトル表現の獲得が可能である。

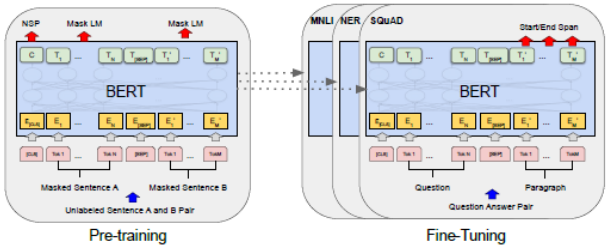

図 2 BERT のモデル構造[2]

他の Transformer を利用したモデルである Elmo や openAI と比較しての BERT の特徴は、文章の学習方向が双方向となった点である。Elmo や openAI では学習方法として左右どちらかへの単方向な単語予測を採用しているが、BERT では双方向での学習が可能な手法が考案された。これにより、特にファインチューニングベースの利用方法での精度が大幅に改善された。

こ BERT の双方向の学習は、マスク単語予測と次センテンス予測と呼ばれる手法で行われる。

BERT の事前学習された言語表現の使用方法としてファインチューニングベース手法と特徴量ベース手法の 2 つがある。

ファインチューニングベース手法の場合には、CLS トークン部分の分散表現が質問応答やセンチメント分析などのタスク分類のために利用可能である。

特徴量ベース手法の場合には、各 Transformer 層を様々な手法で統合し、分散表現を獲得する事によって、BERT 自体を一つの埋め込み層のように利用し、リカレントニューラルネットワークなどへ入力可能である。

3. 使用データ

オープンキャンパス対話システム用データの作成のため、複数人の大学生に受験時の疑問や学内案内で聞きたい質問

の調査を行った。回答は実際に尋ねる際を想定した会話文形式で収集し、それらを内容ごとに 51 の質問クラスへと分類したものをデータとして利用する。

また、分類後各質問クラスについての多様な言い回しを獲得するため、各クラスの内容を尋ねる文章を更に複数人の大学生を対象に調査したものを合わせて使用データする。

質問文と質問クラスの例を表 1 に示す。

表 1　質問文と質問クラスの例

質問文	質問クラス
一般入試について教えて	DS_admission_policy
どうやって大学まで通っていますか	DS_commute
就職状況はどうでしょうか	DS_job-hunting
資格などは取れますか	DS_license

データの総数は 1638 であり、1 クラスには最小 22、最大 47 のデータが属する。

また、質問内容には「授業」クラスに対する「授業難度」クラスのように、より細分化された内容を別の質問クラスとして持つものがあるが、質問は最も適切な分類の単一クラスに属するものとして扱う。

性能比較実験のため、データを質問クラスごとに層別抽出し、学習データ:テストデータ:バリデーションデータ=6:2:2 の割合で分割して使用する。

学習データのみからの言い回しの捕捉度合いとして、分かち書きを行った場合の各データの総単語数と学習データによる出現単語カバー率を表 2 に示す。

表 2　出現単語数とカバー率

	単語数	カバー率
学習データ	685	1
バリデーションデータ	414	0.7609
テストデータ	418	0. 7536

4. 使用モデル

本論文では Word2Vec モデル、BERT ファインチューニングモデル、BERT 特徴量ベースモデル、比較用モデルの 4 つを検討する。

各モデルのパラメータはバリデーションデータに対してのグリッドサーチによって決定した。

4.1 Word2Vec モデル

Word2Vec 利用モデルは、Word2Vec を埋め込み層の重みとして利用し、得られた分散表現を単層の双方向 LSTM へと入力し、双方向 LSTM 層の出力を全結合層で質問分類クラスへ変換し出力する。

学習済み Word2Vec として、東北大学の日本語 Wikipedia エンティティベクトル[3]を使用する。

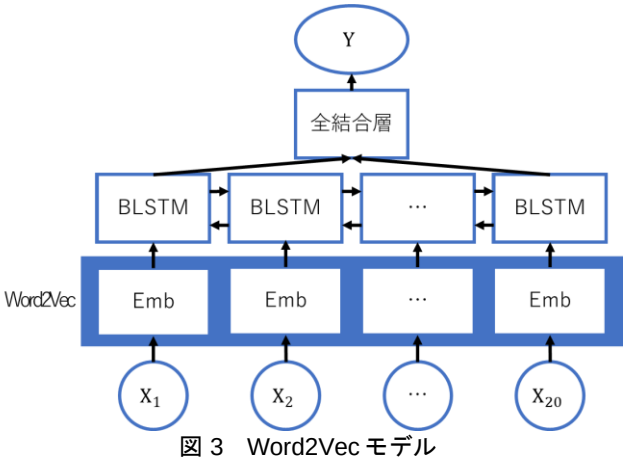

図 3　Word2Vec モデル

モデルの各パラメータを以下に記す。

入力層の受け付ける最大系列長は 20 である。Word2Vec を利用した埋め込み層のユニット数 200 である。双方向 LSTM 層はユニット数 128、双方向のベクトルの連結方法は Concat である。学習はバッチサイズ 64、学習率 0.001、エポック数 100、オプティマイザ adam で行われた。

4.2 BERT 特徴量ベースモデル

BERT 特徴量ベース手法モデルでは、学習済み BERT モデルを単語の出現する文脈を反映した単語分散表現の獲得に使用する手法を検討する。

学習済み BERT モデルには、ファインチューニングモデル・特徴量ベースモデルともに東北大学 BERT[4]を用いる。

BERT のモデル構造からの分散表現抽出手法は定められた手法がなく、Devlin らの論文では 6 つの手法が提案されている。提案法の BERT 特徴量ベース手法モデルでは、BERT 論文で提案されている単語分散表現抽出手法のうち最も CoNLL-2003 Named Entity Recognition でのスコアの高い手法である BERT モデルの Transformer 構造の 9〜12 層の隠れ層を抽出して連結する手法を用いて単語分散表現を獲得する。

抽出した分散表現は単層・双方向 LSTM に入力し、その出力を全結合層で受け取り質問クラスに変換し出力する。

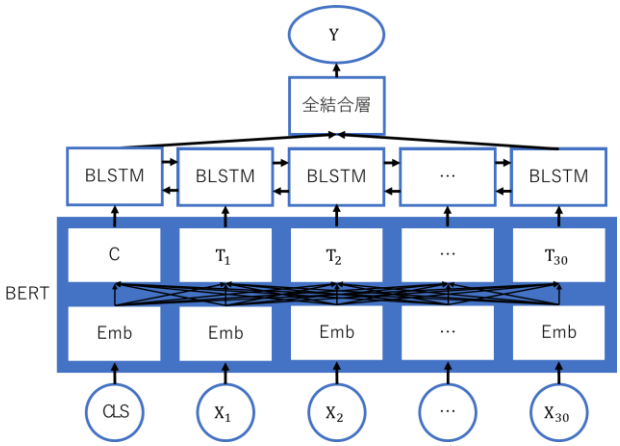

図 4　BERT 特徴量ベースモデル

モデルの各パラメータを以下に記す。
BERT モデル部分は受け付ける入力の最大系列長 30、隠れ層 12、ユニット数 768、HEAD 数 12 である。双方向 LSTM 層はユニット数 256、双方向の隠れ層の出力の連結方法は Concat である。モデル学習はバッチサイズ 64、学習率 0.002、エポック数 20、オプティマイザ Nadam で行われた。

4.3 BERT ファインチューニングモデル

BERT ファインチューニング手法モデルでは、学習済み BERT モデルを質問分類タスクの出力層に直接接続し、ファインチューニングして用いる手法を検討する。
提案法のモデルでは、学習済み BERT の CLS トークン部分が接続された次センテンス予測タスク用の全結合層までを質問分類タスク用の全結合層 1 層に接続し、モデル全体のファインチューニングを行う。

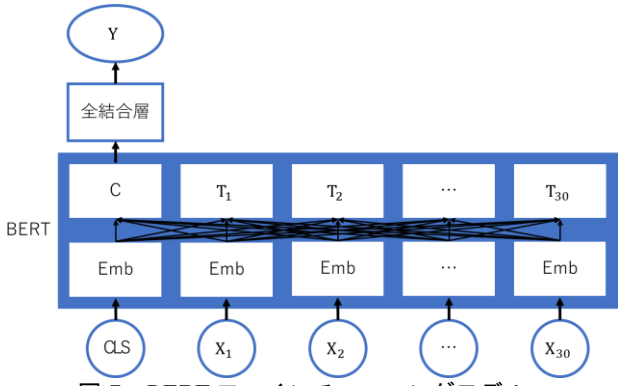

図 5　BERT ファインチューニングモデル

モデル各層のパラメータを以下に記す。
BERT モデル部分は受け付ける入力の最大系列長 30、隠れ層 12、ユニット数 768、HEAD 数 12 である。次センテンス予測タスクの全結合層はユニット数 768 であり、[CLS]に該当する位置のベクトルのみを抽出する。モデル学習はバッチサイズ 64、学習率 0.00005、エポック数 20、オプティマイザ AdamWarmup で行われた。

4.4 事前学習不使用モデル(比較用)

外部データでの事前学習から取得された分散表現の効果を検証するため、事前学習モデルを用いない比較用モデルを構築する。
このモデルではモデル自体の学習とともに、学習データに含まれる単語のみを用いてモデル内で埋め込み層の学習を同時に行う。質問クラス予測時、学習データに含まれない単語は無視される。

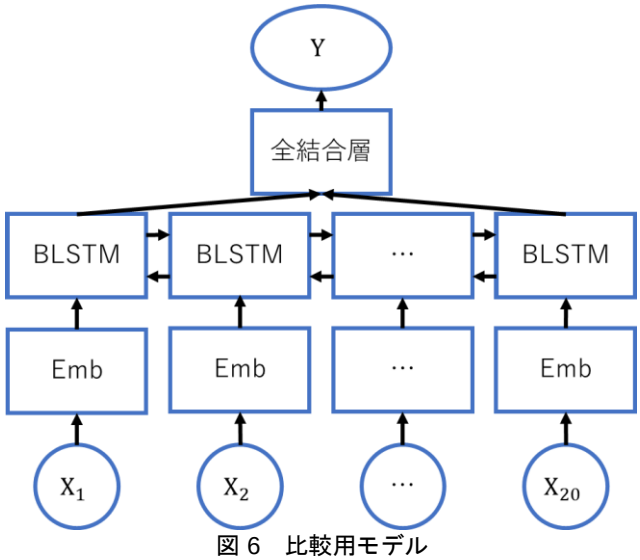

図 6　比較用モデル

モデルの各パラメータを以下に記す。
入力層の受け付ける最大系列長は 20 である。埋め込み層の次元数は 200 である。双方向 LSTM 層はユニット数 128、双方向のベクトルの連結方法は Concat である。学習はバッチサイズ 64、学習率 0.001、エポック数 100、オプティマイザ Adam で行われた。

5. 実験

前章で層別抽出法により全体の 20%を分割したテストデータを用い、各モデルでの予測結果を求めた。
テストデータはクラスごとのデータ数が異なるため、データに偏りのある多クラス分類の評価方法であるマクロ加重平均指標を用いて全体の適合率・再現率・F 値を算出する。
また、GPU によるモデル性能のばらつきを把握するために、本研究では計算に使用される Python、NumPy、TensorFlow のシード値を 0 に固定した上で、GPU 計算によるばらつきの把握方法として Morin[5]らに提案されている複数回の実行結果の分布による性能比較を行う。各モデルでそれぞれ GPU の状態を初期化して 10 回の実験を行い、その平均により性能比較する。

6. 結果

6.1 結果全体の比較

実験により、各モデルのテストデータ全体でのマルチ加重平均適合率、再現率、F 値を比較した結果は表 3 のようになった。

表 3　全体の評価

	適合率	再現率	f 値
分散表現なし	0.8545	0.8345	0.8327
Word2Vec	0.8746	0.8617	0.8604
特徴量ベース	0.8982	0.8953	0.8918
ファインチューニング	0.9259	0.9164	0.9150

テストデータ全体から算出したマクロ加重平均指標による比較では、比較用モデルに対し、Word2Vec 使用モデル、BERT 特徴量ベース手法モデル、BERT ファインチューニング手法モデルの順に性能の向上が見られた。

特に、BERT ファインチューニング手法モデルでは、平均的に F 値 0.91 以上の性能を発揮することができた。

6.2 個別クラスの比較

6.2.1 比較用モデルから性能向上したクラス

比較用の事前学習分散表現をモデルと比べ、提案法モデルにより特に性能が向上したクラスは表 4 のクラスである。

表 4　分散表現により性能向上したモデル

	Word2Vec	ファインチューニング	特徴量ベース	分散表現なし
Shiga-U_mascot	0.7467	1	0.9164	0.6210
Coop_cafeteria_cost	0.8852	0.9364	0.9091	0.6304
DS_lecture_fun	0.7382	0.7576	0.7224	0.5302
Navi_garbage-box	0.9818	0.9727	0.9909	0.7689
DS_gender_ratio	0.8187	0.9909	0.9818	0.8138

これらの質問クラスは、その質問文に高頻度で出現するキーワードとなる単語がないクラスであった。例として「Shiga-U_mascot」では「マスコット」「キャラクター」「キャラ」、「Coop_cafeteria_cost」では「いくら」「どのくらい」「安い」のように近しい意味を表現する際の単語にばらつきが見られた。このため、これらのクラスでは事前学習により類似語が反映された分散表現が特に性能向上に貢献したと考えられる。

6.2.2 比較用モデルから性能低下したクラス

反対に比較用モデルと横ばいの傾向や、使用しない方が性能の良い傾向にあるクラスは表 5 のクラスとなった。

表 5　分散表現により性能低下したモデル

	Word2Vec	ファインチューニング	特徴量ベース	分散表現なし
DS_commute_car	0.9470	0.91212	0.8863	0.9833
DS_graduate_school	0.9596	0.8972	0.8383	0.9667
Shiga-U_bus	0.9535	0.9727	0.9015	0.9909
Shiga-U_domi	0.9909	0.9727	1.0000	0.9909
DS_student_character	0.667	0.7117	0.6012	0.7148
DS_license	0.9900	0.9947	0.9895	0.9947

これらのクラスでは、全質問クラス中でそのクラスにしか登場しない特徴的な単語が質問文に含まれ、他単語に言い換えられることも稀である。例として「車」という単語は「DS_commute_car」、「留学」という単語は「DS_graduate_school」にのみ存在する単語であった。

このため、それらの単語のみに注目するように埋め込み層が学習された比較用モデルの性能の方が良かったと考えられる。

特に提案法モデルの「DS_commute_car」クラスでは、「DS_commute」に分類されるべき「自転車」という単語を含んだ質問文が正しく分類できず、「DS_commute_car」へ分類されてしまうために適合率の低下を招いた。これは、事前学習済み分散表現では「自転車」と「自動車」の類似度が高くなることが原因で分類できなかったと考えられる。

7. 結論

本研究では学部オープンキャンパス用データを例に取り、分散表現技術である Word2Vec や BERT の使用による性能の比較を行った。

実験では、提案法の BERT ファインチューニングによる手法が全体では最も良い性能を発揮し、今回のように比較的少量のデータを用いた分類タスクの性能の向上における BERT を用いたモデルの有用性を確認することができた。

しかし BERT での分類時に性能が低下するクラスがあることも確認できた。

この原因として、分布仮説に基づいた学習済み分散表現では、同じ文脈で使用される単語の類似度が対義語等でも高くなり、人の考える分類とは異なるということが考えられる。そのため、分散表現手法利用時には、それらを考慮した上でのクラス分類等の対策が必要であると考えられる。

謝辞

本研究は科研費（19K02999）の助成を受けた。

参考文献

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality” NIPS'13, Vol.2 pp.3111-3119 (2013).
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre training of Deep Bidirectional Transformers for Language Understanding”, NAACL, pp.4171-4186 (2019).
[3] 東北大学乾研究室, ”日本語 Wikipedia エンティティベクトル” http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector.
[4] 東北大学乾研究室, ”Pretrained Japanese BERT models”, https://github.com/cl-tohoku/bert-japanese.
[5] Miguel Morin, Matthew Willetts, ”Non-Determinism in TensorFlow ResNets” CoRR, abs/2001.11396 (2020).