深層学習による灯謎問題の正解識別システムの構築

1 はじめに

近年深層学習の発展により, 人工知能による漫画や小説などの人間の創作物への理解といった分野の研究が盛んである. 本研究では創作物の一種「クイズ」に注目し, 中国の伝統的クイズゲーム「灯謎 (トウメイ)」を深層学習の手法で正解を識別する方法について提案する.

2 灯謎 (トウメイ)

灯謎は中国の伝統的クイズである. 質問者は問題を詩や熟語の形で出し, 回答者はそれに回答する. 答えは常に字または単語になる. 灯謎は質問に答えるための問題文以外の文書や知識など必要がないものが多く, 質問の文中から答えの情報を得ることが容易である. つまり, 質問を理解すれば回答できると考えられる. 灯謎を解くためには, 問題に隠された情報をもとに, 問われている内容を理解して抽出しなければならないため, 灯謎の研究は一種の情報抽出として考えることもできる.

灯謎のパターンは主に謎とヒントと答えで構成される. 謎は詩や熟語や普通の話し言葉で記述された文である. ヒントは答えの形を説明する文である. ヒントは 1 つ以上与えられ, 答えは字か単語である, 問題に隠された字の構成, 発音, 意味などの情報から解くことができる. 図 1 に灯謎の例を示す.

本研究では灯謎問題のうち,「字謎」と呼ばれる答えが一つの漢字のみとなる種類のみについて考える. 字謎の答えは, 単語の意味に加えて漢字の形も強く関わるので, 単純に大量の問題の文の情報のみをニューラルネットワークで学習しても効果が薄いと予想される. そこで本研究では灯謎の「答え」と「問題」,「ヒント」の関連性形に着目し, 漢字「漢字の画」成分を利用した LSTM モデルで灯謎問題の正解識別システムを構築した.

2.1 中華灯謎データベース

中華灯謎データベース[1]は, 中国各地の灯謎ファン達が集めた灯謎問題 1,362,911 件を収録したデータベースである.

[1]http://www.zhgc.com/mk/

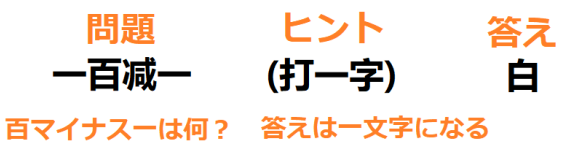

図 1: 灯謎問題

本研究では灯謎のヒントの文を使わないため, 答えが 1 文字である問題 79,725 件のみ利用し, 研究用の灯謎のデータセットを構築した.

研究用灯謎データセットについて, 文中の字の情報のみで解ける問題 72,937 件のみを扱った.

3 要素技術

3.1 SVM

Support Vector Machine(SVM) とは, 分類や回帰などの問題に適用できる機械学習モデルの一つで, データを二クラスに分類する手法である. 特徴として,SVM はカーネル関数を用いることで超平面を生成し, 線形分離が困難な非線形分類問題に適用することもできる. そのため少ないデータ量でも高い精度のモデルを得ることができ, 次元（特徴量の数）が増えても識別精度を維持しやすく, パラメータの調整がしやすいなどの利点がある.

3.2 LSTM

Long short-term memory(LSTM) とは, 深層学習分野に用いられる回帰型ニューラルネットワーク構築モデルの一種類である. 特徴として,LSTM のセルはデータを連続的処理することができ, 長時間にたってその隠れ状態を維持することができる. 故に回帰型ニューラルネットワークに起きる勾配消失や勾配爆発などの問題を解決できる.

3.3 Word2Vec

Word2vec は,Mikolov らが 2013 に提案したモデルである. 二層のニューラルネットワークのみで構成され

るという特徴により, モデルの計算量は比較的少なくなり, 大規模なデータで分散表現を学習することが可能となる.

3.4 BERT

Bidirectional Encoder Representations from Transformers(BERT) は,2018 年に Google が公開した自然言語モデルである.Transformer による双方向エンコーダ構造を利用し, 文脈を考えることが可能になり, 多様なタスクにおいて高い表現を得られる.

3.5 Levenshtein 距離

Levenshtein 距離は,2 つの文字列がどの程度異なっているかを示す距離の一種である. 具体的には,1 つの文字列からもう一方の文字列に変形するのに必要な手順 (1 文字の挿入, 削除, 置換) の最小回数を計算する.

4 提案手法

4.1 二分類灯謎データセットの作成

漢字の「形的情報」が灯謎に対する有効性を検証するために, 二分類灯謎データセットを生成する. 具体的には, 研究用灯謎データセットの「問題; ヒント; 正解」テキストを基づいて,「問題; ヒント; 不正解」のデータを生成し, そして「問題; ヒント; 正解」テキストに「True」ラベルを付け,「問題; ヒント; 不正解」テキストに「False」ラベルを付けていた.

生成手法について,「正解のデータに見たことない漢字が出る」に起こすデータ不均衡問題を解決するため, 本研究は正解の漢字を基に, Levenshtein 距離が 1 である漢字で「問題; ヒント; 不正解」のデータを生成する. そのため, 同じ漢字に対して,「正解」と「不正解」として扱われる回数は同じになり, データ不均衡問題が実験結果に与える影響を最小化にさせた.

4.2 画の分散表現

象形文字の一つとして, 漢字は絵文字からの発展によって生まれたと考えられている. しかし従来の研究は中国語の意味的情報を重視するものが多い, 常に漢字を最小単位として扱われている. そして 2021 年には Chen らが漢字の「画」を「横棒」,「縦棒」,「左払い」,

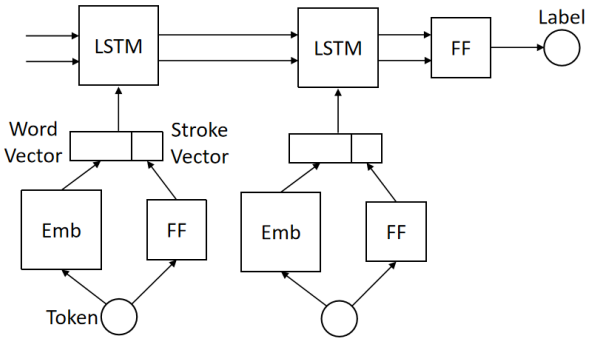

図 2: 提案手法

「点」,「鉤」の 5 種類に分類し, 書き順で漢字を分解する手法を提案した [?].

本研究は Chen らの手法に基づき, 漢字の「画」の種類により 5 次元ベクトルを生成し, そして全結合層で分散表現を生成する手法を提案し, そして灯謎問題の正解識別モデルに使用した.

4.3 灯謎問題の正解識別モデルの構築

本研究は LSTM と Word2Vec を利用し, 加え漢字の画情報の灯謎問題の正解識別モデルを提案した. embedding 層について, まず入力データを漢字に分け,Word2Vec で漢字の意味的情報を含まれた分散表現を生成し, そして同じ漢字に対して, 漢字の 5 次元画ベクトルを生成し, そして全結合層で画の分散表現を生成した.

次に,Word2Vec で生成した分散表現と画の分散表現を結合し,LSTM に入力してラベルを生成する.

図 2 にモデル構造を示す.

5 実験

5.0.1 データ処理

本実験では中華灯謎データベースから収集した「問題; ヒント; 正解」テキスト 72937 件と提案手法で生成した「問題; ヒント; 不正解」テキスト 72937 件, 合計 145874 件のデータの中に, 出現頻度 (正解として扱われる回数) が 3 以下の漢字を削除しデータセットを作成した. そしてその中からランダムに訓練データ 106464 件, テストデータ 26616 件を抽出し, 実験に使用した.

表 1 に実験用データを示す.

表 1: 資料のデータ数

訓練データ	テストデータ
106464	26616

表 2: 実験用パラメータ (LSTM)

パラメータ	数値
分散表現の次元数	300
画の分散表現の次元数	8
隠れ層の次元数	256
バッチサイズ	128
Dropout	0.5
最適化手法	Adam
学習率	0.001
Epoch	200

5.1 実験内容

実験では訓練データとテストデータに対し,「Tfidf ＋ LinearSVM」,「LSTM」,「Word2Vec ＋ LSTM」,「Word2Vec ＋画分散表現＋ LSTM」,「BERT」の条件で実験を実施し, 訓練誤差とテスト精度を計算した.

表 2 に実験モデルのパラメータを示す.

5.2 実験結果

200 Epoch を経た訓練結果として,「Word2Vec ＋画分散表現＋ LSTM」の条件でのテスト精度は 0.805 に収束し,「Word2Vec ＋ LSTM」条件より良い効果 (精度は 0.792) があることが確認した.

図 3 と図 4 に提案手法の訓練誤差とアキュラシィ変化を示す.

図 5 と図 6 に Word2Vec を用いた LSTM モデルの訓練誤差とアキュラシィ変化を示す.

図 7 と図 8 に LSTM モデルの訓練誤差とアキュラシィ変化を示す.

図 9 と図 10 に BERT の訓練誤差とアキュラシィ変化を示す.

表 3 に訓練結果を示す.

表 3: 各モデルの実験結果

結果	SVM	LSTM	Word2Vec ＋ LSTM	Word2Vec ＋画分散表現＋ LSTM	bert
訓練誤差		0.249	0.122	0.132	0.030
訓練精度	0.543	0.880	0.951	0.945	0.982
テスト精度	0.371	0.762	0.792	0.805	0.613

テスト結果について,「BERT」モデルは訓練データ

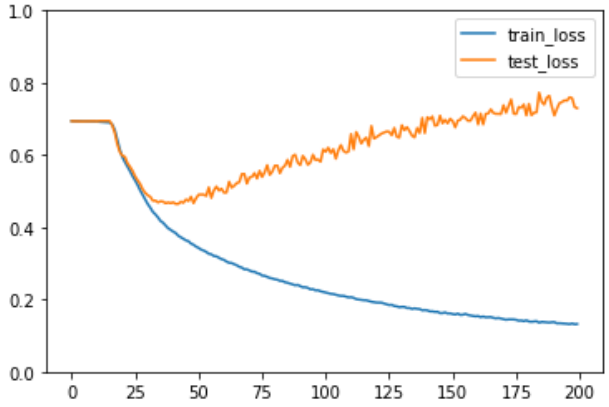

図 3: 訓練誤差曲線 (提案手法)

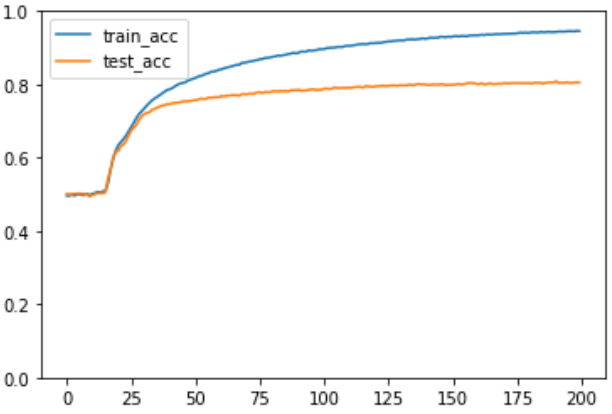

図 4: アキュラシィ(提案手法)

に対して最も良い表現を得られるが, テストデータに対する表現は「LSTM」モデルの方が高い, その中に, 提案手法の精度が最も高いため, 漢字「画像的情報」は一定程度に有効することを確認した.

6 まとめと今後の課題

本研究は Word2Vec と LSTM を用いて, 漢字の画情報を導入した灯謎問題の正解識別モデルを提案した. 結果として, 提案手法は既存手法より良い効果があることが確認した.

今後の課題として, 漢字の画情報を BERT などのモデルに導入し, 正解識別の精度向上についてもふれておく.

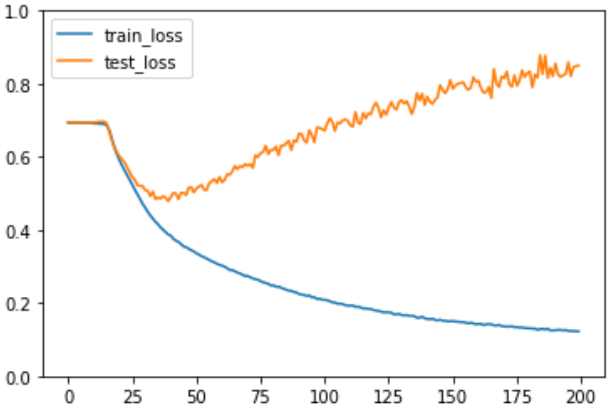

図 5: 訓練誤差曲線 (Word2Vec + LSTM)

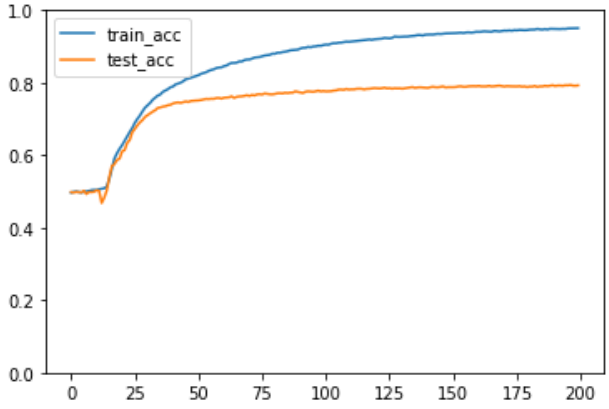

図 6: アキュラシィ (Word2Vec + LSTM)

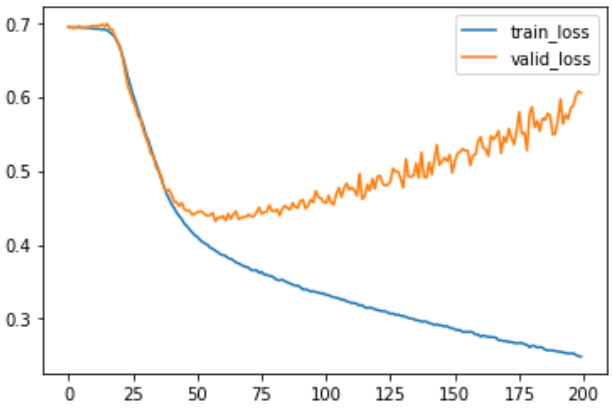

図 7: 訓練誤差曲線 (LSTM)

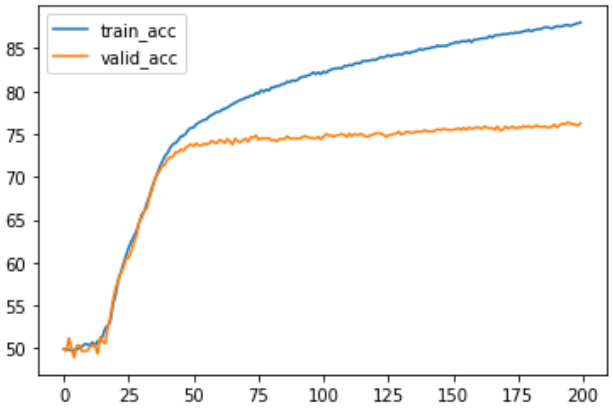

図 8: アキュラシィ (LSTM)

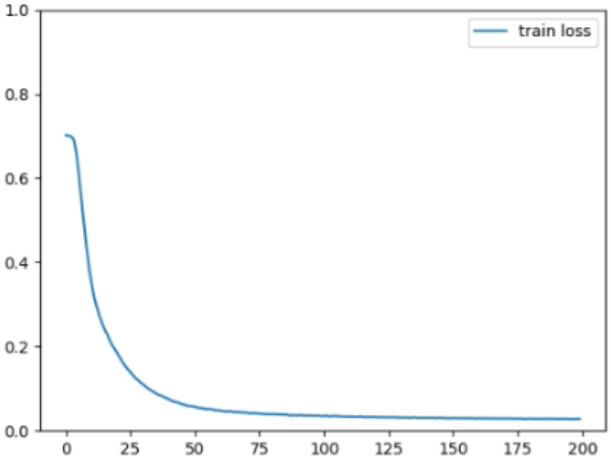

図 9: 訓練誤差曲線 (BERT)

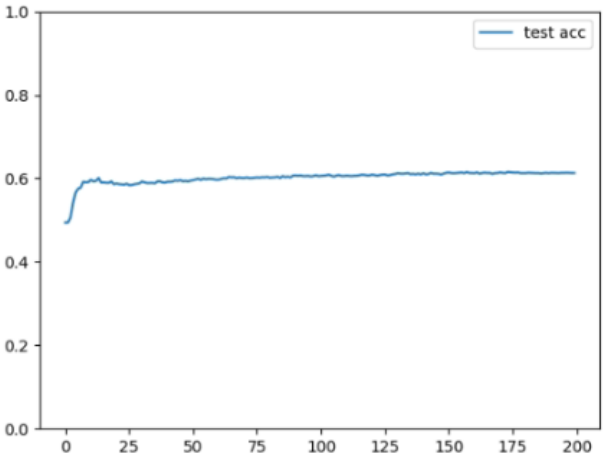

図 10: アキュラシィ (BERT)