深層学習による灯謎問題の正解推定システムの構築

第 1 グループ 陳 偉斉

1. はじめに

深層学習の発展により, 人工知能による漫画や小説やクイズゲーム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である「クイズ」に注目し, 中国の伝統的クイズ「灯謎 (トウメイ)」を深層学習の手法で正解を推定する方法について提案する.

2. 灯謎

「灯謎 (トウメイ)」は中国の伝統的クイズである. 図 1 に灯謎の例を示す. 質問者は問題に当たる「謎面」とヒントに当たる「謎目」を作成し, 紙で灯篭に張り, 回答者は答えに当たる「謎底」を当てる. 長文の中のキーワードを探すような読解問題と違い, 灯謎は質問に答えるための問題文以外の長文や知識など必要がないものが多く, 質問とヒントのみで答えの情報を得ることができる. そのため, 灯謎の解答に必要な情報は,「謎面」を構成する漢字の構造的情報と意味的情報となる. この点から灯謎は一種の情報抽出として考えることもできる.

「謎面」は詩や熟語などで記述された短い文であり,「謎目」は答えのパターンを説明するヒントである. どんな年齢層でも楽しめるように, 灯謎の「謎底」は常に字または単語であり,「謎面」に隠された字の構成, 発音, 意味などの情報に加えて「謎目」のヒントで解くことができる.

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「謎底 (答え)」と「謎面 (問題)」,「謎目 (ヒント)」の関連性に着目し, 漢字の「画」成分を利用した LSTM モデルで灯謎問題の正解推定システムを構築した.

3. データセット

3.1. 中華灯謎データベース

中華灯謎データベース[1] は 2002 年に灯謎愛好者が公開した灯謎問題データベースである. 現時点では中国各地の灯謎問題 1,404,526 件が収録されている. 収録された灯謎は 9 種類に分けられており,「謎面」,「謎目」,「謎底」,「作者」,「備考」の形式で保存されている. 本研究はその中の「字謎」と呼ばれる答えが 1 文字である灯謎 79,725 件を収集し,「謎面の字の情報のみで解ける字謎」72,937 件のみ使用した.

3.2. Levenshtein 距離

Levenshtein 距離 [1] 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.

3.3. 二値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, 二値分類灯謎データセットを生成する. 具体的には, まずデータセットの中にある低頻度語 (出現頻度 4 以下) を除き, 次に字謎データセットの「問題; ヒント; 正解」を「正例」として「True」ラベルを付け, そして「正解の漢字の画」と Levenshtein 距離で「簡単な不正解 (Levenshtein 距離が大きい)」を生成し,「同じ問題; 同じヒント; 不正解」を「負例」として「False」ラベルを付ける. 最後にデータセットを 8 対 2 の比率で訓練データとテストに分けて実験する.

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

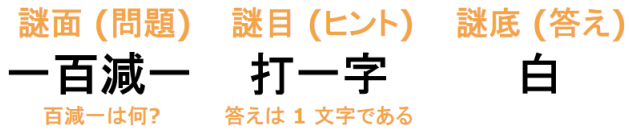

図 1: 灯謎の例

表 1: データセットの情報

	訓練データ	テストデータ
データ総数	106464	26616
正例	53277	13263
負例	53187	13353

データ不均衡問題を解決するため, 正解と不正解のデータ数をそろえた. これによ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する. 表 1 にデータセットの情報を示す.

4. 提案手法

4.1. 画ベクトルの生成

従来の研究では中国語の意味的情報を重視するものが多く, 漢字を最小単位として扱っている研究が一般的である. 2021 年に Chen らは五筆字型入力方法 [2] を用いて, 漢字の「画」を「横棒」,「縦棒」,「左払い」,「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 [3].

本研究は Chen らの手法に基づき, 上記の 5 種類の漢字の「画」で漢字の構造情報を表現し, 5 次元ベクトルで漢字の画ベクトルを生成することで漢字の構造情報を分類器に追加した. 図 2 に漢字の画の分散表現ベクトル生成例を示す.

4.2. 灯謎問題の正解推定分類器の構築

本研究は Long short-term memory (LSTM)[4] と Word2Vec[5] を利用し, 漢字の画情報を加えた灯謎問題の正解識別モデルを提案する. embedding 層では, まず入力データを漢字に分割し, Word2Vec で漢字の分散表現ベクトルを生成する. そして同じ漢字に対して漢字の画ベクトルを生成し, 全結合層で画の分散表現ベクトルを生成する.

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

図 3 に提案モデル構造を示す.

5. 数値実験

5.1. 実験設定

提案手法で示した正解推定分類器を用いて, データセットにおける灯謎問題の答えが正解か否かを 2 値分類で予測する. 実験のハイパーパラメータは Optuna によって調節し, 過学習問題による影響を最小化させる. そして提案手法を従来の分類手法である機械学習モデル, LSTM, LSTM (Word2Vec), BERT と比較する. 表 2 に実験モデルのパラメータを示す.

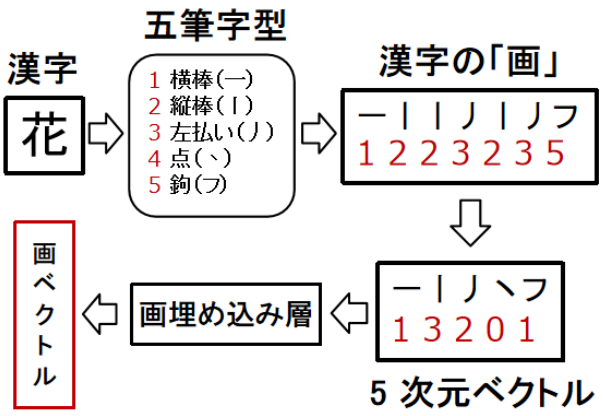

図 2: 画ベクトルの生成

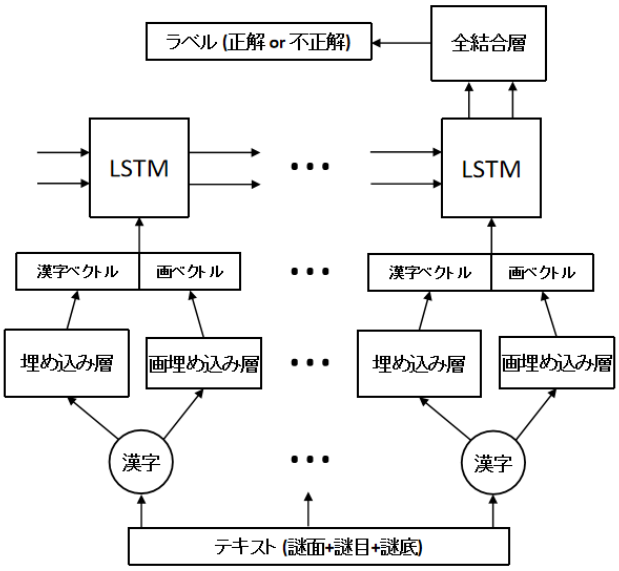

図 3: 提案モデルの構造

5.2. 実験結果

表 3 に各手法による灯謎の正解の予測結果を示す. “LSTM+Word2Vec+Stroke” は提案手法である. 実験結果は精度を示す Accuracy と F1 値で評価した.

結果として, 提案手法である “LSTM+Word2Vec+Stroke” を用いた場合の Accuracy と F1 値がそれぞれ最大となり, 漢字の画情報の特徴量は灯謎正解を推定するのに有効であることが示せた. 図 4.a にテストデータの例と説明を示せし, 図 4.b に各手法における結果の比較を示す.

6. まとめと今後の課題

本研究では中国の伝統的クイズゲーム「灯謎」のデータセットを新たに構築し, 漢字の画情報を利用した手法で灯謎問題の正解推定をした.

今後の課題として, 答えが 1 文字の「字謎」に加えて, 全部の灯謎問題に利用できる手法の考案,「画」より詳細な漢字の構造的情報を含まれた「SUB 漢字」を導入した灯謎問題の正解推定の精度向上が挙げられる.

表 2: 実験用パラメータ

パラメータ	数値
バッチサイズ	128
Dropout	0.5
最適化手法	Adam
学習率	3e-5
Epoch	200

表 3: 各手法によるテストデータの予測評価値

手法	Accuracy	F1 値
SVM+Tfidf	36.8	36.5
Random Forest+Tfidf	50.1	16.3
MLP+Tfidf	49.4	30.5
BERT	73.3	73.2
LSTM	80.4	80.9
LSTM+Word2Vec	81.7	82.2
LSTM+Word2Vec+Stroke (提案手法)	85.7	85.7

	謎面(問題)	謎目(ヒント)	謎底(答え)
正例	差点損国格	四筆字	王(四筆)
負例	差点損国格	四筆字	树(九筆)
翻訳	ほぼの背信行為	答えは四画の漢字	
説明	“差点” は漢字 “王” の点を除く, “損国格” は漢字 “国” の罷を除くを意味するため正解は漢字の “王” である		

(a) テストデータの例と説明

method	Label	Prediction
SVM+Tfidf	FALSE	TRUE
Random Forest+Tfidf	FALSE	TRUE
MLP+Tfidf	FALSE	TRUE
BERT	FALSE	TRUE
LSTM	FALSE	TRUE
LSTM+Word2Vec	FALSE	TRUE
LSTM+Word2Vec+Stroke	FALSE	FALSE

(b) 各手法における結果の比較

図 4: テストデータの例の説明および結果の比較

参考文献

[1] V. I. Levenshtein, et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10, pp. 707–710. Soviet Union, 1966.

[2] L. Surhone, M. Tennoe, and S. Henssonow. *Wubi Method, Wubizixing Input Method*. VDM Publishing, 2011.

[3] J. Chen, B. Li, and X. Xue. Zero-shot chinese character recognition with stroke-level decomposition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 615–621, 2021.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality, 2013.