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)」正解を深層学習の手法で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, 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タスク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 Long short-term memory (LSTM)[1] モデルで灯謎問題の正解推定システムを構築した.

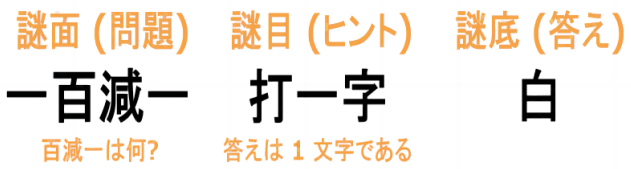

図 1: 灯謎の例

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	106,464	26,616
正例	53,277	13,263
負例	53,187	13,353

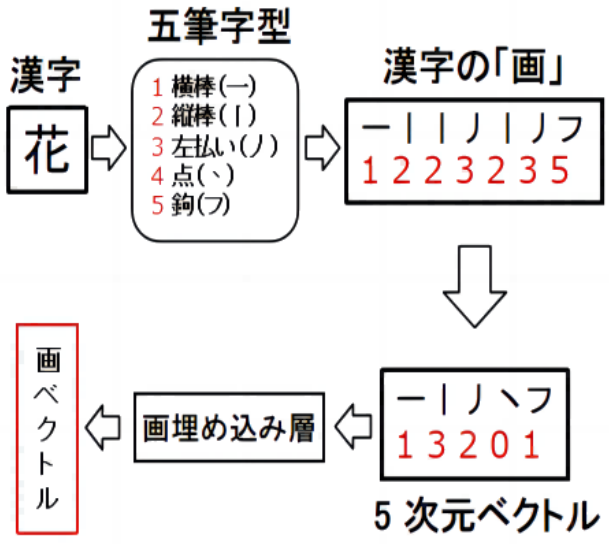

図 2: 画ベクトルの生成

3. 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.

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.

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

4.2. Levenshtein 距離

Levenshtein 距離 [4] 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) の最小編集回数を計算する. 本研究は Levenshtein 距離を用いて他の漢字から正解に変形する最小編集回数を計算することで正解の漢字と似ていない漢字 (Levenshtein 距離が大きい漢字) を生成し, 不正解として研究用二値分類灯謎データセットを作成する.

4.3. 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作成する. 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

表 2: 各手法によるテストデータの予測結果

手法	Accuracy	precision	Recall	F1 値
BERT	73.26%	74.08%	72.40%	73.22%
LSTM+Word2Vec	81.68%	79.73%	84.78%	82.18%
LSTM+Word2Vec+Stroke (提案手法)	85.76%	85.55%	85.94%	85.74%

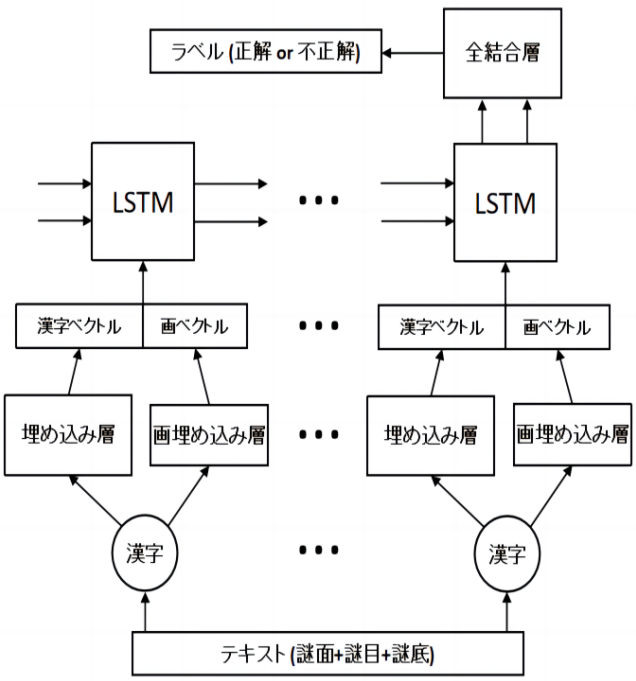

図 3: 提案手法によるモデル構造

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 (似ていない漢字)」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.

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揃えた. そのため同じ漢字に対して正解として扱われる確率は不正解と等しくなり, データ不均衡問題の影響を最小にする. 表 1 にデータセットの情報を示す.

4.4. 灯謎問題の正解推定分類器の構築

本研究は LSTM 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 に提案手法によるモデル構造を示す.

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二値分類で予測する. 実験のハイパーパラメータは Optuna によって最適化する. そして提案手法を従来の分類手法である LSTM (Word2Vec) および BERT[6] と比較する.

表 3: 提案手法で正解となった正解数

手法	問題数
BERT	3,316
LSTM+Word2Vec	1,087

実験モデルのパラメータ設定として, 分散表現の次元数は 300, 画の分散表現の次元数は 30, 隠れ層の次元数は 256, バッチサイズは 128, Dropout は 0.5, 最適化手法は Adam, 学習率は 0.00003, Epoch 数は 200 とした.

5.2. 実験結果

表 2 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, Precision, Recall, F1 値で評価した. 表 3 に従来手法では不正解だったが, 提案手法では正解となった問題数を示す.

提案手法により Accuracy, Precision, Recall, F1 値がそれぞれ最大となった. これにより, 漢字の画情報の特徴量が灯謎の正解を推定することに有効であることが示した.

6. まとめと今後の課題

本研究では中国の伝統的クイズ「灯謎」のデータセットを新たに構築し, 漢字の画ベクトルを用いた灯謎問題の正解推定の手法を提案した. 漢字の画ベクトルを導入することで, 灯謎問題の正解推定モデルは漢字の構造情報を利用でき, 従来手法を上回る精度を達成した.

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情報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.

参考文献

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

[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] 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.

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

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, June 2019.