人工知能による灯謎問題解決システムの構築

1 はじめに

質問応答タスクは文書をもとに入力された質問に対して正しく応答することを目的とするタスクである. 今まで質問応答に対する研究は読解問題と呼ばれ, 主に問題と答えの情報を含む長文を対象としてきた. その一方で, 問題の中に隠された情報で問題を解けるクイズ問題と呼ばれタスクも存在する. 本研究ではこのクイズ問題にとりくむ. 具体的には人工知能で中国の伝統的クイズ問題「灯謎 (トウメイ)」を解く方法について考える.

2 灯謎 (トウメイ)

灯謎は, 中国の伝統的クイズ問題である. 質問者は問題を詩や熟語の形で出し, 回答者はそれぞれ回答する. 答えは常に字または単語になる. 質問応答とは違い, 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く, 質問の文中から答えの情報を得る. 言い換えと, 質問を理解すれば回答できる. 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ので, 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.

灯謎のパターンはだいたい, 謎とヒントと答えで構成される. 謎は詩や熟語, あるいは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 つ以上与えられる答えは字か単語であり問題に隠された字の構成, 発音, 意味などの情報からで解くことができる. 図 1 に灯謎の一つの例を示す.

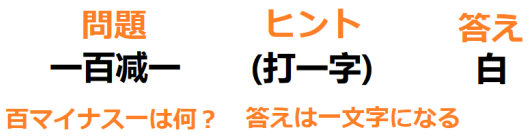


図 1: 灯謎の例

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意味がない. そこで本研究では漢字の形的情報に着目し, 漢字と SUB 漢字成分を利用した Sequence to Sequence モデルで灯謎問題解決システムを構築した.

3 要素技術

3.1 Sequence to Sequence

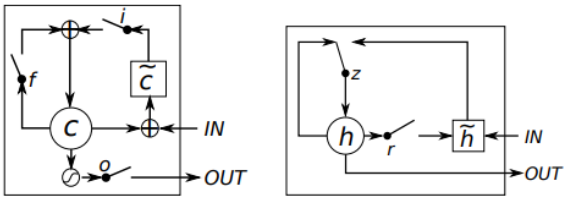
Sequence To Sequence (Seq2Seq) [?] とは, 2014 年に Google が発表した言語モデルである. 従来の Deep Neural Network (DNN) が扱いにくい時系列データ問題を解決するため, seq2seq は Encoder-Decoder という形式のモデル構造を導入した. Encoder は入力する時系列データをベクトルに圧縮し, そのベクトルを Decoder に渡し出力系列を生成し, 機械翻訳タスクによく使われているモデルである.

3.2 Gated recurrent unit

Recurrent Neural Network (RNN) [?] とは, 回帰構造を持つニューラルネットワークである. 通常のニューラルネットワークでは, レイヤの出力は次のレイヤの入力として利用されるが, RNN では同じレイヤーに対して現時刻の時系列データだけでなく, 前時刻の出力も合わせて入力する.

誤差逆伝播法による RNN の訓練は, 逆伝播される勾配の消失 (勾配がゼロに収束) , あるいは爆発 (勾配が無限に発散) する問題がある. この問題を解決するため, ゼップ・ホッフライターらは 1997 年に Long short-term memory (LSTM) [?] を提唱した. LSTM のアーキテクチャは Memory Cell と三つの Gate (Input Gate, Output Gate, Forget Gate) から構成される. LSTM は勾配をそのまま使用することが可能であるので, 勾配消失と勾配爆発の問題を解決できる.

Gated Recurrent Unit [?] とは, LSTM の Input Gate と Forget Gate を統合して隠れ状態のみ利用するニューラルネットワークの状態である. 図 2 に LSTM と GRU の構造を示す.



(a) Long Short-Term Memory (b) Gated Recurrent Unit

図 2: LSTM と GRU の構造 [?]

4 データセット

4.1 漢字の分け方

漢字の分け方について, 主に大まかに分ける Shallow Mode と細かいに分ける Deep Mode 二つの分け方がある.

図 3 に漢字の分け方を示す.

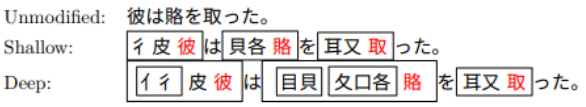


図 3: 漢字の分け方 [?]

4.2 IDS データセット

IDS(Ideographic Description Sequence)[?] とは, 中国語, 日本語, 韓国語の漢字データを 'unicode', ' 漢字', ' サブ漢字' の形で集まったデータセットである.

本研究の実験対象灯謎は中国語で作られたため,IDS の中国語データを利用し, IDS のない漢字の ' 画' 情報を手動で補足した.

そして, IDS の中に漢字成分不足という問題 (漢字「爽」の締める部分など) に対して, 漢字の画の書き順で補足した.

表 1 に本研究に使われている漢字数と手動で SUB 漢字成分を補充した数に示す.

表 1: 使用した漢字数と補充したデータ数

使用した漢字数	補充した漢字数	画を補充した漢字数
9285	782	9285

4.3 中華灯謎ベース

中華灯謎ベース [?] とは, 中国各地の灯謎ファン達か集まった灯謎 1,362,911 問が収録した.

本研究では灯謎のヒントを使わないため, 答えが一文字である灯謎問題を利用し, 灯謎のデータセットを構築した.

灯謎データセットについて, まずは「数式で解決できる問題」は「答えの漢字の成分は問題の中 (漢字或いは漢字の成分) にある」と定義した.

以上の定義により, 最初に集まった灯謎 79725 問を数式問題と非数式問題に分け, 算式問題のみ利用した.

表 2 に問題の類を示す.

表 2: 問題の類

問題の総数	数式問題	非数式問題
79725	72937	6788

5 提案手法

分類モデルで灯謎問題を解くと, 答えである漢字の数が 8648 あるため, モデルの精確率はとても低いである. そこで本研究は, 出力を漢字の「画」とする Seq2Seq による灯謎問題の解決方法を提案した.

Cao らの研究 [?] により, 漢字の「画」は五つの類に分類できる. この手法と Seq2Seq モデルを利用すると, モデルの出力は 8648 から 9 次元に削減できる.

5.1 モデル構造

本研究のモデルでは, まず問題を漢字に分け, 順次で Embedding 層で分散表現に変換し, そして Encoder に入力して問題の情報を含まれた Vector を生成した.

続いて, 問題の情報を含まれた Vector と「開始」を表示した「EOS」信号を Decoder に入力し, 答えは SUB 漢字を書き順で出力した.

図 4 に実験に用いた seq2seq モデルの構造を示す.

6 実験

6.1 データ処理

本実験では実験用データを 8:1:1 の比率で Train Data, Validation Data, Test Data に分け, 実験に使用した.

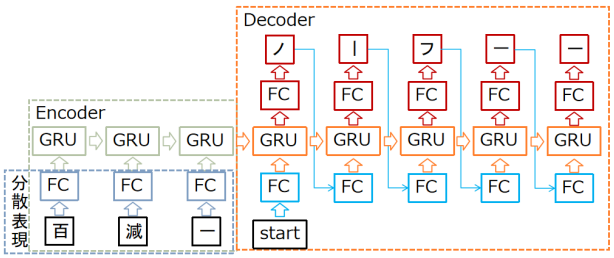


図 4: モデルの構造

表 3 に実験用データ数を示す.

表 3: 実験用データ数

Train Data	Validation Data	Test Data
58350	7293	7294

6.2 実験内容

本研究は漢字の成分を利用する可能性を確認するため, 入力する問題の分かち書き方法により, 「Word」, 「Word + Char」, 「Word + Char + Shallow」, 「Word + Char + Deep」, 「Word + Char + Shallow + Deep」, 「Char」, 「Char + Shallow」, 「Char + Deep」, 「Char + Shallow + Deep」九つの状況で, 各自 GRU を基づいた Seq2Seq モデルで対照実験を行っていた.

表 4 に各実験の入力次元数を示す.

表 4: 各実験入力次元数

実験	入力次元数
Word	51783
Word + Char	53507
Word + Char + Shallow	53897
Word + Char + Deep	53701
Word + Char + Shallow + Deep	58350
Char	4561
Char + Shallow	4951
Char + Deep	4755
Char + Shallow + Deep	4971

表 5 にその他のパラメータを示す.

表 5: 実験用パラメータ (通用)

パラメータ	数値
出力サイズ	9
分散表現の次元数	256
隠れ層の次元数	512
バッチサイズ	128
Dropout	0.5
最適化手法	Adam
学習率	0.001
損失関数	Cross-Entropy

6.3 実験結果

100 epoch の訓練を経て, 各実験の Train Loss と Valid Loss は表 6 に示す.

表 6: Train Loss と Valid Loss

実験	Train Loss	Valid Loss
Word	0.158	3.905
Word + Char	0.193	3.725
Word + Char + Shallow	0.152	3.748
Word + Char + Deep	0.228	0.535
Word + Char + Shallow + Deep	0.219	3.528
Char	0.688	2.318
Char + Shallow	0.460	2.623
Char + Deep	0.455	2.653
Char + Shallow + Deep	0.489	2.611

図 5, 図 6 に各誤差の変化曲線を示す.

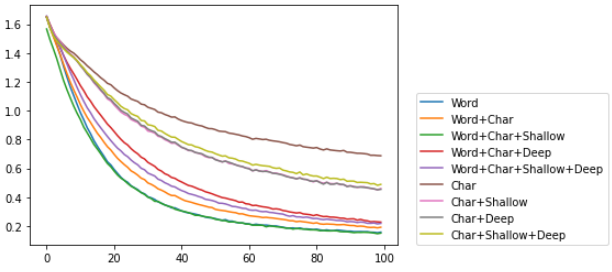


図 5: Train Loss の変化曲線

Accuracy は,Valid Data と Test Data による「Precision」,「Recall」,「F1」の値で計算した.

表 7 に Valid Data の Accuracy を示す.

図 7, 図 8, 図 9 に各誤差の変化曲線を示す.

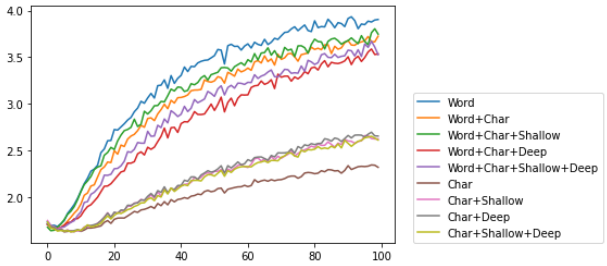


図 6: Valid Loss の変化曲線

表 7: Valid Accuracy

実験	Precision	Recall	F1
Word	0.685	0.754	0.685
Word + Char	0.677	0.765	0.684
Word + Char + Shallow	0.709	0.752	0.697
Word + Char + Deep	0.679	0.751	0.679
Word + Char + Shallow + Deep	0.678	0.766	0.683
Char	0.628	0.759	0.652
Char + Shallow	0.673	0.740	0.669
Char + Deep	0.645	0.753	0.659
Char + Shallow + Deep	0.638	0.759	0.657

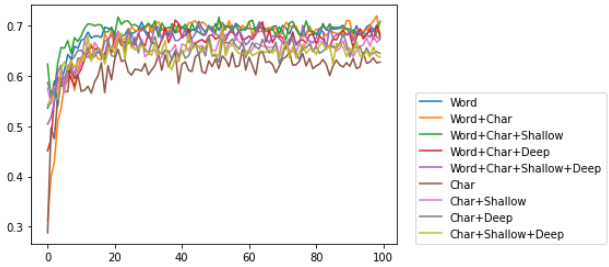


図 7: Valid Precision の変化曲線

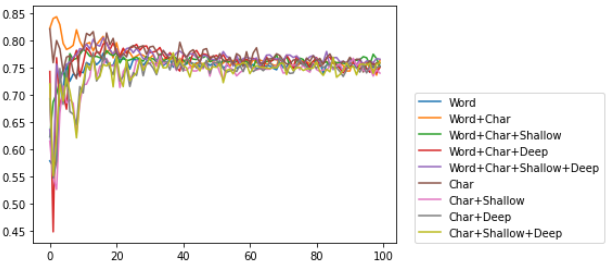


図 8: Valid Recall の変化曲線

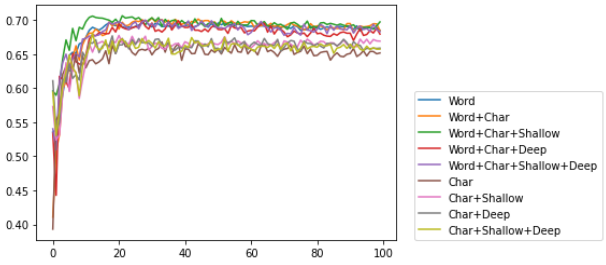


図 9: Valid F1 の変化曲線

表 8 に Test Data の Accuracy を示す.

表 8: Test Data Accuracy

実験	Precision	Recall	F1
Word	0.541	0.713	0.592
Word + Char	0.506	0.824	0.595
Word + Char + Shallow	0.553	0.686	0.589
Word + Char + Deep	0.597	0.676	0.610
Word + Char + Shallow + Deep	0.609	0.724	0.636
Char	0.576	0.771	0.631
Char + Shallow	0.625	0.686	0.625
Char + Deep	0.633	0.665	0.620
Char + Shallow + Deep	0.633	0.672	0.624

7　まとめと今後の課題

本研究は, GRU による Seq2Seq を基づいて, 灯謎問題モデルを構築し, 各入力条件の対照実験で精度を確認した. 結果として, 全実験の精度は低いだが,「Word + Char + Shallow + Deep」という条件の精度は比較的に高いことを確認した.

今後の課題として, Attention Mechanism を利用した新しいモデルと IDS 漢字構造情報の導入で現有モデルの精度向上を目指す. そして,「問題」を画像で解決するモデル, 或いは「漢字」から「問題」を生成するモデルの構築を試す

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[2] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, Vol. 404, p. 132306, Mar 2020.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[4] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, Vol. abs/1412.3555, , 2014.

[5] Viet Nguyen, Julian Brooke, and Timothy Baldwin. Sub-character neural language modelling in Japanese. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 148–153, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[6] `https://github.com/cjkvi/cjkvi-ids`.

[7] `http://www.zhgc.com/mk/`.

[8] Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. cw2vec: Learning chinese word embeddings with stroke n-gram information. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018.