人工知能による灯謎問題の解決手法の調査 (仮)

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.

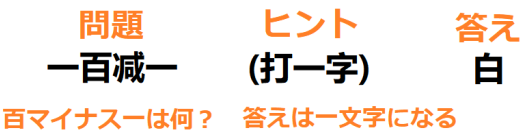

図 1: 灯謎の例

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 モデルで灯謎問題の答えを生成する.

3　要素技術

3.1　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. 図 2 に RNN の構造を示す.

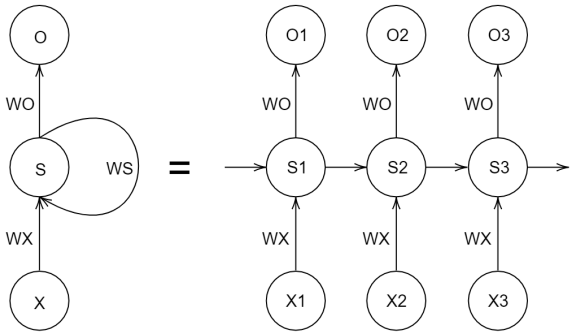

図 2: RNN の構造

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. 図 3 に LSTM の構造を示す.

Gated recurrent unit とは, 2014 年に公開された LSTM の簡易版である.

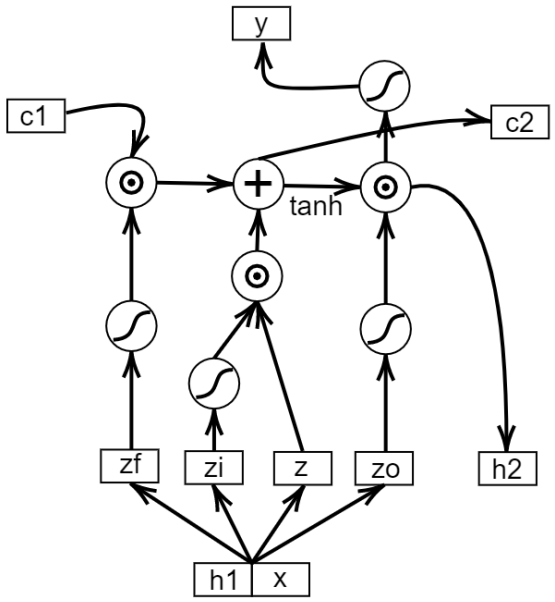

図 3: LSTM の構造

3.2 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する. 本実験の seq2seq モデルは RNN を利用したため, Decoder の出力は自動的に調整される.

4 データセット

4.1 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」, 「画」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 の中国語データしか使わない.

4.2 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.

5 実験

5.1 データ処理

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 4 に漢字の分け方を示す.

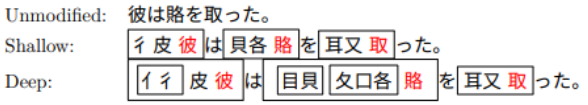

図 4: 漢字の分け方 [?]

今回実験は IDS の Shallow Mode を利用したが, ある SUB 漢字は表現できない問題 (例えば漢字 ' 爽' 中の締める部分は IDS データセットで表現できない) がある.

故にこれらの表現できない部分を手動でより細かい画にわけ, 書き順で補充した.

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
9285	782

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 57535 問と非数式問題 22190 問に分け, 算式問題だけ利用した.

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	57535	22190

最後に実験のデータを 8:1:1 の比率で Train Data, Validation Data, Test Data に分け, 実験に使用した.

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
46029	5753	5754

5.2 モデル構造

本研究は漢字から SUB 漢字, SUB 漢字から SUB 漢字二つの状況により, 各自 GRU を基づいた Seq2Seq モデルと Attention Mechanism を利用した GRU を基づいた Seq2Seq モデルで対照実験を行っていた.

具体的に, 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「start」信号を Decoder に入力し, 答えは SUB 漢字を書き順で出力した.

図 5 に実験に用いた seq2seq モデルの構造を示す.

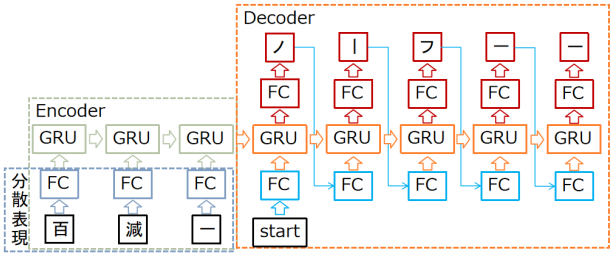

図 5: モデルの構造

表 4 に漢字から SUB 漢字実験のパラメータを示す.

表 4: Char2Sub モデルのパラメータ

パラメータ	数値
入力サイズ	4312
出力サイズ	1518
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
損失関数	Cross-Entropy

表 5 に SUB 漢字から SUB 漢字実験のパラメータを示す.

表 5: Sub2Sub モデルのパラメータ

パラメータ	数値
入力サイズ	1230
出力サイズ	1518
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
損失関数	Cross-Entropy

5.3 実験結果

10 epoch の訓練を経て, 漢字から SUB 漢字の場合, GRU を基づいた Seq2Seq モデル (Char2Sub_GRU) の Train 誤差と Validation 誤差は 1.775 と 2.883 に収束し, Attention Mechanism を利用したモデル (Char2Sub_Attn) の Train 誤差と Validation 誤差は 0.295 と 3.658 に収束した. 図 6, 図 7 各モデルの誤差変化曲線を示す.

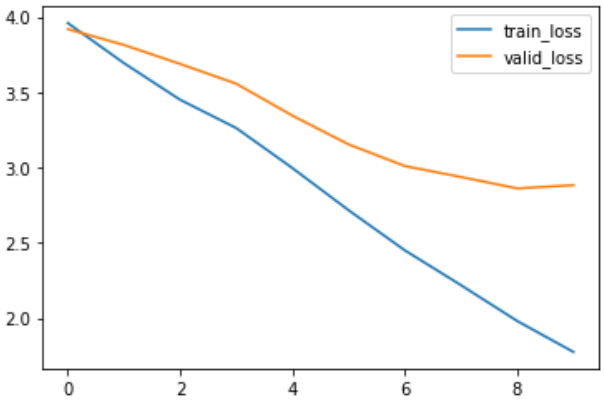

図 6: Char2Sub_GRU 誤差曲線

SUB 漢字から SUB 漢字の場合, GRU を基づいた Seq2Seq モデル (Sub2Sub_GRU) の Train 誤差と Validation 誤差は 1.766 と 2.902 に収束し, Attention Mechanism を利用したモデル (Sub2Sub_Attn) の Train 誤差と Validation 誤差は 0.582 と 3.208 に収束した. 図 8, 図 9 に各モデルの誤差変化曲線を示す.

Accuracy は, Test Data による「予測結果のあってる個数」/「正解の個数」で計算した.

表 6 に各実験の結果を示す.

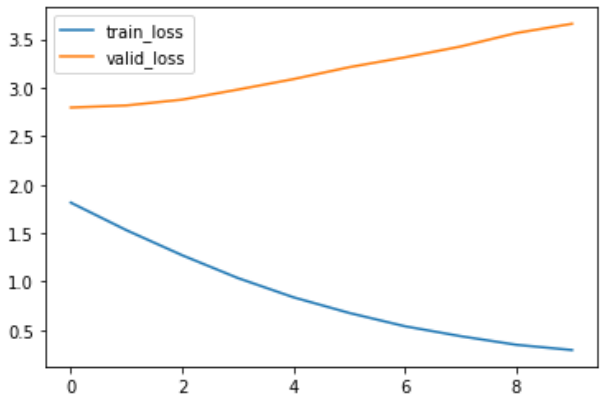

図 7: Char2Sub_Attn 誤差曲線

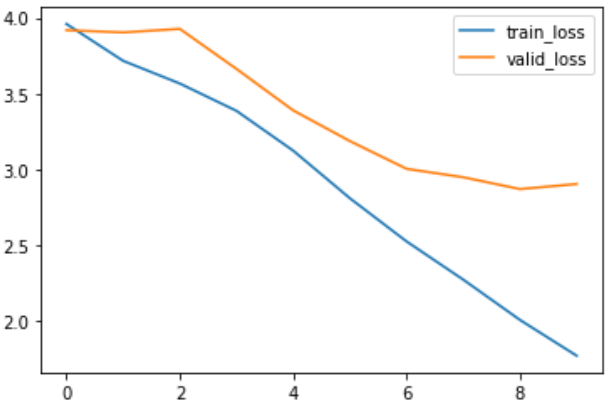

図 8: Sub2Sub_GRU 誤差曲線

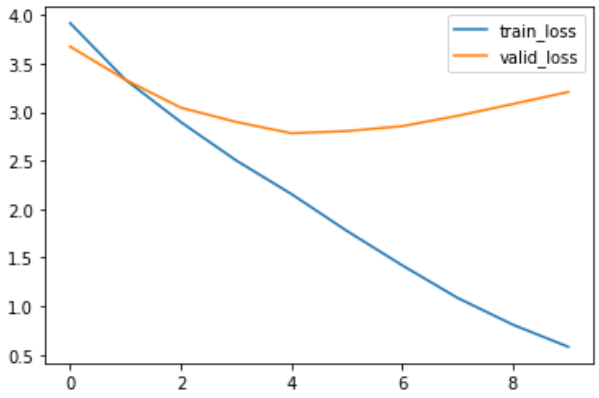

図 9: Sub2Sub_Attn 誤差曲線

表 6: 各実験の結果

実験	Char2Sub_GRU	Char2Sub_Attn	Sub2Sub_GRU	Sub2Sub_Attn
Train Loss	1.775	0.295	1.766	0.582
Validation Loss	2.883	3.658	2.902	3.208
Test Accuracy	0.261	0.302	0.246	0.288

6　今後の課題

データについて, Seq2Seq を利用したため, 答えは一文字に限らない方が, 中華灯謎ベースのデータをより多く利用でき, そして灯謎問題の「ヒント」部分も利用できる.

モデルについて, Transformer と Bert を利用して実験すること (作成中).

Accuracy の計算について, まだ改善する点がある.