進捗報告

1 はじめに

質問応答タスクとは, 文書をもとに, 入力された質問を正しく応答することを目指すタスクである. そして対象とする知識の範囲が限らない質問応答はオープンドメイン質問応答と呼ばれる. 今まで質問応答に対する研究は主に問題と答えの情報を含める長文で問題を解ける (読解問題). しかし, 問題の中に隠された情報で問題を解けるタスクも存在する (クイズ問題). 本研究ではクイズ問題に着目し, 人口知能で中国の伝統的クイズ問題灯謎を解ける方法を探索する.

2 灯謎 (トウメイ)

灯謎とは, 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提出し, 回答者が回答する. 答えは常に字や, 単語の形式である. 質問応答とは違い, 灯謎は質問を回答するための文書や知識など必要はない, 単に質問の中に答えの情報を得る, 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, 謎に隠された情報をもとに, 質問を理解しなければならないので, 灯謎の研究は人間の言葉の情報抽出として扱うができる.

灯謎のパターンはだいたい, 謎, ヒント (時にはヒント 2 もある) と答えで構成される. 謎は詩や熟語, あるいは普通のしゃべり言葉で作る. ヒントは答えの形を説明する. 答えは問題に隠された情報 (字の構成, 発音, 意味など) で解くことができる.

解決方法について, 灯謎問題を字謎 (答えは一つの漢字のみ) のみに考えると, マルチラベル問題にも考えられる. しかし, 字謎の答えは, 単語の意味と漢字の形に関わるので, 簡単に大量な問題をニューラルネットワークに通すでも意味がない. そこで本研究漢字の形に着目し, 漢字の部首情報を利用し,CharCBOW と CharSkipGramで問題文を処理する.

3 要素技術

3.1 分散表現

分散表現（あるいは単語埋め込み）とは, 単語を高次元の実数ベクトルで表現する技術である. 機械は簡単に人間の言葉などの自然言語を理解することができないため, 自然言語を分散表現の形変換する必要性がある.

3.2 Word2Vec

Word2Vec とは,Mikolov が 2013 に提案した, 単語の分散表現を生成するためのアルゴリズムである.Word2Vec は,2 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がある. この特徴により, 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は可能になる.Word2Vec には CBOW 法と SkipGram 法二つの手法が含めている.

CBOW は前後の単語から目標単語を予測する手法である.CBOW 法では, 入力として周辺語を与え, その中心語の予測を出力する. この学習を通じて, ネットワークにある単語の周囲に, どのような単語が現れる可能性が高いのかを学習させる.

CBOW と違い,SkipGram は目標単語から前後の単語を予測する手法である.CBOW の場合, 入力は中心語, 出力は周辺語となる.

3.3 charCBOW と charSkipGram

今まで分散表現に関する研究は, 分散表現の粒度では語, 句や文に限らない. 特に中国語に関する研究は主に単語や漢字を文脈の最小単位として扱っている. しかし, 漢字の中で隠された情報に注目する研究は少ないである. 漢字は, 部首と偏旁で構成されている, この中には漢字の意味を提示する情報も含めている. 例えば, 漢字花と草の部首はくさやはなみたいな植物の意味が含めている. この部分情報は質問応答や感情分析などのタスクに利用する可能性がある.CharCBOW と CharSkipGra mは Yanran らが 2015 に提案した, 漢字と漢字の成分の漢字埋め込み方法である.Word2Vec の手法に基づき,CharCBOW の入力は周辺語の漢字と漢字の成分情報の結合ベクトルである. そして結合ベクトルを入力とし, 中間語を予測する.CharSkipGra mは CharCBOW と違い, 中間語から周辺語の漢字結合ベクトルを予測する形である.

4　データセット

本研究は三つのデータセットを使用する. 分散表現を訓練するため,zhwiki で漢字に分けで実験する. そして zhwiki とオンライン新華辞書を利用し,zhwiki と対応する zhwiki　 radical データセットを構築する.

灯謎の実験に対して, 個人作者に灯謎を収集し, 灯謎のデータセットを構築した.

5　実験

5.1　データ処理

a

5.2　モデルの実装

a

5.3　トレーニング

a

5.4　実験結果

a

6　まとめと今後の課題

- 漢字の画像と音声情報の処理手法の探索
- 灯謎生成問題に対する取り組み