令和 3 年 10 月 27 日 (水) ゼミ資料 M1 陳偉斉

進捗報告

1 今週やったこと

- 灯謎問題を解決するための取り組み

2 灯謎問題を解決するための取り組み

取集された 200 字謎 (答えは一文字) 問題を分析した結果, 漢字の形で解決する方が多いので, 灯謎問題を解決するため, 漢字の形の情報抽出問題を解決しなければならないである. そしてニューラルネットワークは漢字の形を理解する機能がないので, 単に問題を入力すると良い結果が出る可能性は低い.

漢字の形の情報抽出問題の解決法は二つの視点から考える.Su ら [1] は漢字を画像として扱い, そして新しい中国語単語の埋め込み方法を提案した.

もう一つは漢字の中に情報を抽出すること. ほうこう従来の研究は単語, あるいは漢字を中国語文脈の最小単位として扱うことが多いが, 漢字の成分に関する研究は少ないである. 英単語は語源で構成せれるように, 漢字も部首と偏旁で構成される, その部首と偏旁の中には漢字の意味に関する情報が含めっている. 故に漢字を偏旁部首に分け, そしてニューラルネットワークに入力することが可能になる.

方法二と違い,Li ら [2] は,CBOW と SkipGram に基づいて, 漢字の偏旁部首情報を含める埋め込み方法 CharCBOW と CharSkipGram を提案した. この方法に参考し, 漢字と偏旁部首だけでなく, 単語情報も含める単語埋め込みモデルの構築は可能になる.

これからも方法二について実験する.

3 実験設計

現有の灯謎問題データセットの量は少ないので, 灯謎問題データセットで単語埋め込みモデルを訓練することは難しいである. 故に Sougou Copus[3] で訓練したいと思う. そして漢字の偏旁部首はオンライン新華辞書で抽出することができる. 漢字から漢字と偏旁部首の埋め込み表現に変換するので,CharSkipGram に参考したモデルの方がいいと思う. 最後は LSTM などのネットワークで字謎データセットを実験する. 字謎の答えは全部一文字, そして漢字の形の情報に注目させるため, 字謎のヒント 2 (答えはどういうものを提示する情報) を全部考えなしで実験する.

4 来週目標

- 漢字埋め込みモデルを訓練する
- 字謎をできるだけ取集

参考文献

[1] Tzu-Ray Su and Hung-yi Lee. Learning chinese word representations from glyphs of characters. *CoRR*, Vol. abs/1708.04755, , 2017.

[2] Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. Component-enhanced Chinese character embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 829–834, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3] Yiqun Liu, Fei Chen, Weize Kong, Huijia Yu, Min Zhang, Shaoping Ma, and Liyun Ru. Identifying web spam with the wisdom of the crowds. *ACM Trans. Web*, Vol. 6, No. 1, March 2012.