

## 進捗報告

### 1 今週やったこと

- 研究会からの問題のまとめ

### 2 問題のまとめ

- 灯謎問題をパターンに分け, 算式で解ける問題と算式で解けない問題に分けて実験すること
- 日本語の漢字の分散表現に対して, データセットを利用する可能性を確認すること
- 漢字の分散表現と部首が含まれた漢字の分散表現のみならず, 部首だけの分散表現も対象として実験すること
- 漢字の画を利用すること
- 直接に灯謎データセットで分散表現を生成すること

### 3 灯謎のパターン

灯謎は算式で解ける問題と算式で解けない問題に分類されてます. この中に算式で解ける灯謎を収集して実験することは可能である.

### 4 日本語漢字の分散表現データセット

「文字には有益な情報が詰め込まれている !?」 [1] という論文から見ると, 日本語の文字と sub-文字に対するデータセットは GlyphWiki, IDS, KanjiVG, KRADFILE 四つが存在している. これらのデータセットの資料を見ると, 中国語の研究に利用するか, それとも中国語バージョンを作るかを考える.

### 5 部首の分散表現

現有のプログラムで部首だけの分散表現を作ることも可能になる. 学習したモデルで生成した分散表現が「汗 - 水 + 月 = 肝」みたいな関係があれば, 算式で解ける灯謎に使うことも可能になる.

### 6 漢字の画を利用した研究

2018 年に Cao らが, 漢字の画を利用した cw2vec[2] を提出した. これから論文を読み, 理解する.

### 7 灯謎データセットで分散表現を生成する

一定の量な灯謎が集まれば, 実験する.

## 8 来週目標

- 発表資料を訂正する
- 論文を読む

## 参考文献

- [1] <https://ai-scholar.tech/articles/treatise/text-ai-181>.
- [2] 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.