進捗報告

1　今週やったこと

- 論文内容のまとめ
- 灯謎問題の収集

2　中国漢字に関する研究

2.1　Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components

論文は中国漢字の部首と部首以外の成分などのサブ漢字成分に注目し,CBOW 構造を用いた単語 + 単語構成する漢字 + 漢字を構成するサブ漢字で分散表現を生成する JWE[1] モデル提案した. 図 1 にモデルの構造を示す.

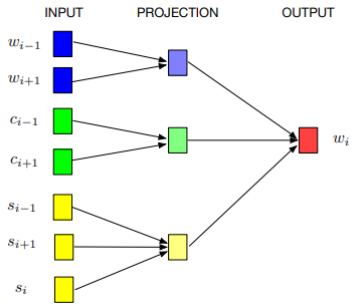


図 1: JWE[1] の構造

2.2　cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information

論文は中国漢字の画に注目し, 単語を漢字の画の n-gram 特徴にで表現し, そして SkipGram 構造のモデルで前後文を予測する cw2vec[2] モデルを提案した.

図 2 にモデルの構造と単語画の n-gram 特徴を示す.

3　日本漢字に関する研究

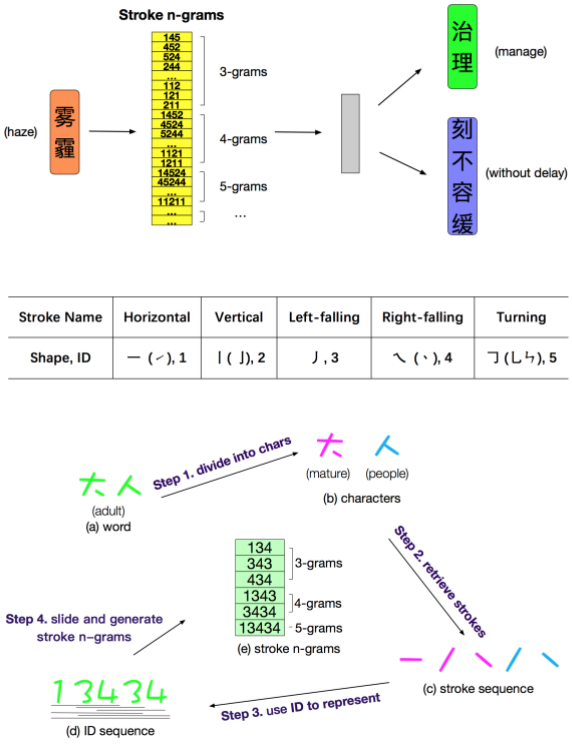
3.1　Sub-character Neural Language Modeling in Japanese

論文は日本漢字を構成するサブ漢字を対象として, 四つの漢字データで実験した [3]. 図 3 に「賂」を例として, データセットそれぞれの分解方法を示す.

著者らは NAIST コーパスを利用し, モデルは単向 LSTM で次のトークンを予測する形になる.

漢字分解の程度により、モデルに入力したテキストは図 4 に示したように, 浅い分解と深い分解にそれぞれ実験をする.

結果として, IDS 漢字データセットの浅い分解は一番いい結果が出る.



Stroke Name	Horizontal	Vertical	Left-falling	Right-falling	Turning
Shape, ID	一 (㇀), 1	丨(亅), 2	丿, 3	乀 (㇏), 4	乛(㇄), 5

図 2: cw2vec[2] の構造

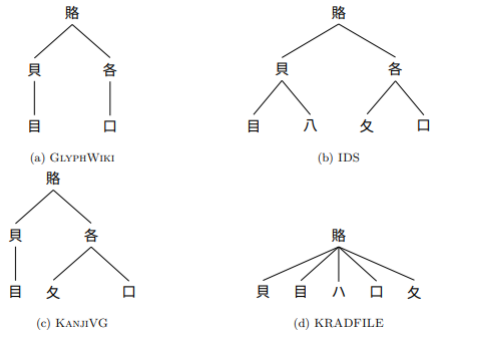


図 3: データセットの例 [3]

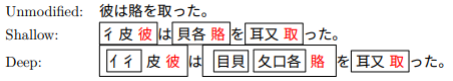


図 4: 漢字分解の例 [3]

3.2 Subcharacter Information in Japanese Embeddings: When Is It Worth It?

論文はサブ漢字単位を対象とした分散表現が日本語に使う可能性を確認するため，IDS データセットでの浅い分解で単語類似らのタスクで実験した.

書者らは SkipGram モデルを利用し, 単語 + 単語構成する漢字 + 漢字を構成するサブ漢字で分散表現を生成する SG+kanji+bushu[4] モデルを構築した.

図 5 に SG+kanji+bushu モデルを示す.

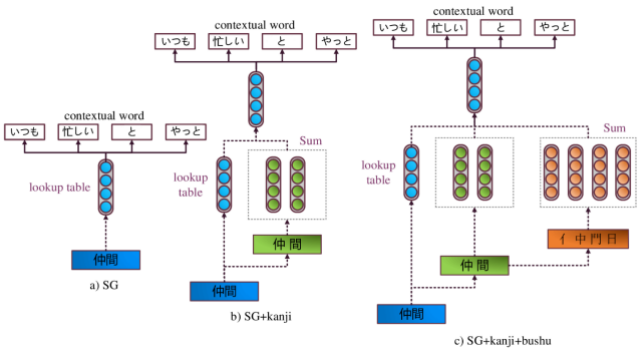


図 5: SG+kanji+bushu[4] の構造

そして結果として, 中国語にいい表現があるサブ漢字の分散表現の利用は, 日本語には漢字の多いテキストの方がいい表現があると確認した.

4 灯謎問題の収集

中華灯謎データベースに 130 万くらいの灯謎が集まっている. その中に, 答えは字である灯謎は少なくとも 10 万以上ある.

これらのデータを人工で抽出するには時間かかりすぎであるため, ソフトウェアで抽出する.

5 来週目標

- 灯謎問題データセットの完成と分析

参考文献

[1] Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 286–291, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[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.

[3] 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.

[4] Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. Subcharacter information in Japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pp. 28–37, Melbourne, Australia, July 2018. Association for Computational Linguistics.