Contextual-based Chinese Word Segmentation and Word Sense Disambiguation

卢雨轩 19071125

2022 年 6 月 10 日

Table of Contents

Introduction

Task

- Chinese Word Segmentation
- Measure: Precision, Recall and F1 score
  - GOLD: **共同 创造 美好 的 新 世纪 ——二○○一年 新年 贺词**
  - OUTPUT: **共同 创造 美 好 的 新 世纪 ——二○○一年 新年 贺词**
  - Precision: $\frac{TP}{PP} = \frac{10}{11} = 0.909$
  - Recall: $\frac{TP}{P} = \frac{10}{10} = 1$
  - F1: $\frac{2*P*R}{P+R} = 0.952$ **调和平均**

Task

- Word Sense Disambiguation
  - 小米手机就是好用
  - 我今天吃了一碗小米粥
  - 牙膏我只用中华为的就是刷的干净速度快，每次只用 5G

Related Works

Related Works

BERT

BERT

- BERT[1]: Bidirectional Transformer for Language Understanding
- Transformer Encoder, Attention
- Pre-training and fine-tuning
  - Pretrained Language Model
  - Mine contextualized semantic information
  - Achieved SOTA on multiple downstream tasks

Related Works

MetaSeg

MetaSeg

- MetaSeg[2]: Pre-training with Meta Learning for CWS
- BERT-Based
- Meta Learning
  - Learning from multiple datasets
  - Learn difference of datasets
    - CTB6: **李娜/进入/半决赛**
    - PKU: **李/娜/进入/半/决赛**
    - MSRA: **李娜/进入/半/决赛**
  - Put dataset label into BERT

Method

CWS

- Use BERT to obtain contextualized embedding
  - Word embedding
  - Position embedding
  - Contextual Semantic Information
- WordPiece?
- Use a simple MLP to decide whether to segment
- Minimise NLL Loss

WSD

- Use BERT to obtain contextualized embedding
- k-MEANS
- t-SNE

Experiment

Experiment Environment

- AMD Ryzen 9 5950X
- GEFORCE 3090
- PyTorch 1.10
- One epoch

Dataset

- SIGHANS 2005 Bakeoff dataset
- MSR and PKU

数据集	字数	训练集	划分训练集	划分验证集	测试集
MSR	4050566	86924	80000	6924	3985
PKU	1826475	19056	18000	1056	1945

表 1: 数据集信息

Result

Entropy

- on MSR dataset
- Character Entropy: 9.43 bit
- Word Entropy: 11.1 bit

模型	PKU	MSRA
BERT(Yang, 2019)	96.50	98.40
MetaSeg(Ke,2021)	**96.92**	**98.50**
BERT(ours)	95.60	98.40

表 2: 汉语分词任务试验结果

- 『美』
  - 美丽，美好，美妙 ...
    - “一个人不管学什么专业，总得懂一些文学知识，有一点艺术素养，这对于丰富自己的思想和生活，提高自己的审美能力很有好处。
    - “传说再美丽再动听，终归是传说。
  - 美国，中美，欧美，美元 ...
    - 二战期间，中美两国是同盟国成员，在战场上并肩战斗。
    - 影片的编剧和导演是美国人，由中美两国著名演员主演。

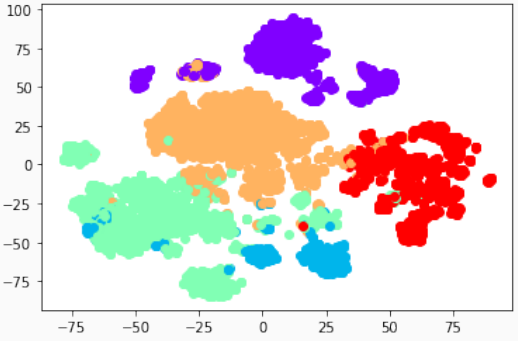

图 1: 聚类结果

References I

[1] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186. DOI: 10.18653/v1/N19-1423.

[2] KE Z, SHI L, SUN S, et al. Pre-training with Meta Learning for Chinese Word Segmentation[C/OL]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 5514-5523. DOI: 10.18653/v1/2021.naacl-main.436.