進捗報告

1 今週やったこと

- 中国語日本語文脈の分かち書き処理
- プログラムの実行

2 中国語日本語文脈の分かち書き処理

英語文脈は、単語を最小単位として処理されています。単語の間に空白があるので、処理は中国語や日本語より簡単である。中国語と日本語の場合、単語の間に区別用な空白がないので、実験する前に、まずは分かち書きの予処理を行う。

今週の実験は主に、ManyThings の ENG-CMN データと ASPEC-JC データセットを回って実験している. 中国語 データの処理は、jieba で行う. 日本語データの処理は、janome で行う. そしてプログラムの実行をスピートアップさせるため、データは ASPEC-JC のテストデータで実行する. 表 1 に各データセットの文のペア数を示す. 表 2 表 3 に各データセットの言語ごとの単語数を示す.

表 1: Pairs

DataSet	ManyThings	ASPEC-JP
Pairs	24360	2107

表 2: ManyThings_Words

ManyThings	ENG	CMN
Words	7484	14716

表 3: ASPEC-JC_Words

ASPEC-JP	JPN	CMN
Words	6640	6844

単語データはディクショナリの形で表示されている。図1図2に各データセット言語ごとのindex2words形式の例を示す。

Input index2word CMN - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

文件(F) 編輯(E) 播試(O) 書書(O) 器制(H) (O: 'SOS, 1: 'EOS, 2: '確', 3: 'a. '('NG')', 5: '('R', 6: '用', 7: ')證, 8: '('B')', 9: '等等', 10: '! '! ', 11: '等', 12: '一下', 13: '开始', 14: '我', 15: '高', 16: '\了', 17: '不会', 18: ''吧', 19: '*整杯', 20: '', 21: '!', 22: '懂', 23: '\"', 24: '\"', 25: '\"', 26: '\"', 26: '\"', 26: '\"', 27: '\"', 26: '\"', 27: '\"', 28: '\", 30: '\"', 31: '\"', 29: '\"', 31: '\"', 22: '\"', 21: '\"', 25: '\"', 26: '\"', 26: '\"', 27: '\"', 28: '\"' ,下雪",204: "3",205: "兵",206: '天",207: '食物",208: '秦胄",208: "秦胄",208: "春",208: "秦胄",208: "秦胄",208: "秦胄",208: 船', 255: '我錯', 256: '要,257: '守时, 258: '鳥兒', 259: '歌唱', 260: '可以', 261: '幫忙', 262: '?', 263: '必须', 264: '不要', 265: '动', 266: '', 267: '等待', 268: '填滿', 269: '限', 270: '著', 271: '洗洗', 272: 上帝', 273: '存在', 274: '让步', 275: '凶', 276: '很穷', 277: '嘿', 278: '放松', 279: '工作', 280: '怎么样', 281: '赶快', 282: '个头', 283: '矮', 284: '我会', 285: '做饭', 286: '我會', 287: '不能', 288: '這裡', 289. '嫉妒', 290: '羡慕', 291: '恨', 292: '认识', 292: '弘宗', 294: '喜欢', 295: '喜歌', 297: '囊', 298: '寒', 299: '鬼念', 300: '需要", 301: '惶用', 302: '這樣', 303: '使用', 304: '「雪', 304: '雪', 307: '秦', 296: '喜歌', 307: '秦', 296: '喜歌', 308: '常月', 317: '他长', 318: '那', 319: '爱嗎', 320: '真是', 321: '太好子', 322: '晚上', 323: '安 静', 324: '想一想', 325: '祈祷', 326: '走走', 327: '那里', 328: '禁止', 329: '评论', 330: '无可奉告', 331: '没', 332: '问题', 333: '再', 334: '一次', 335: '请用', 336: '座', 337: '往后', 338: '始', 339: '保持警 静,324: '地是一想,325: '祈祷,326: '未走,327: '那里,326: '秦止,329: '评论',330: '无可奉告',331: '没了,332: '问题',333: '再,334: '一次,335: '请用,336: '座,337: '往后,338: '站,339: '保持警惕,340: '留在,341: '这里,342: '让开,343: '好好学习',344: '真疼,345: '畸光',346: '滞逝,347: '匆匆,348: '走过',349: '失败',350: '打命欠',351: '向右转,352: '失去,353: '男孩,354: '吊宫,355: '什么',356: '有,357: '新鲜事,358: '端里,359: '端上,360: '下,361: '木村,362: '燃煤,363: '完成,364: '生气,365: '爆心',366: '友好',367: '秦子,368: '强一,369: '境,370: '荥,370: '荥,370: '荥,370: '荥,370: '荥,370: '荥,370: '荥,370: '荥,370: '荥,370: '大四,370: '荥,370: '荥,370: '荥,370: '太阳,370: 541: '真糟糕', 542: '学校', 543: '欣赏', 544: '几乎', 545: '祈禱', 546: '蘋栗', 547: '魚子', 548: '酱', 549: '我給', 550: '留下', 551: '不了', 552: '无所谓', 553: '一辆车', 554: '一条', 555: '狗', 556: '一張',

図 1: 中国語 index2words の例

III output index2word JPN - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

裳, 368: '本く, 369: 'ほか、370: '更素、371: 'つ, 372: 'なか、373: 'り, 374: '明確、375: '必要、376: '上, 377: '則し、378: '本業、379: 'マネシメント'、380: '環局、381: '重要、382: '視点、383: '本稿、384: '規格、385: '、, 386: '環, 387: '間線、388: '取り、389: 'け、390: '高。391: '汚染、392: '製薬、393: '廃。394: '拍響、395: '浓。396: '湿式、397: '空気、380: '腹化、399: 'による'、400: '予備、401: '处理、402: '発酵、403: '降散、404: '本日、409: 'ロット'、410: '食荷、411: '十分、412: '低た'、413: 'ない、414: '本日、401: 'なり、412: '低た'、413: 'ない、414: '本日、401: 'なり、412: '低た'、413: '太い、414: '本日、401: 'なり、412: '低た'、413: '太い、414: '本日、401: 'なり、412: '低た'、413: '太い、414: '本日、412: '伝た'、413: '太い、414: '本日、412: '伝た'、413: '太い、414: '本日、412: '伝た'、413: '太い、414: '本日、412: '伝た'、413: '太い、414: '太い、 第1行,第1列

図 2: 日本語 index2words の例

実験を行う 3

今回の実験は、データを seq2seq モデルに入力し、75000 エポックに実行する. 図 3 図 4 に各データセットの実験誤 差を示す. 実験結果として,ManyThings の ENG-CMN データを 30 分ほどのトレーニングした後, 平均誤差は 2.7713 になる. そして,ASPEC-JC データは,40000 エポックを超えると, 平均誤差が 3.5 から 16.4 に急増な状況がある.

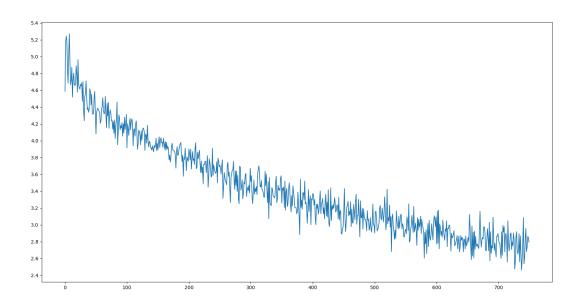


図 3: ManyThings 実験誤差

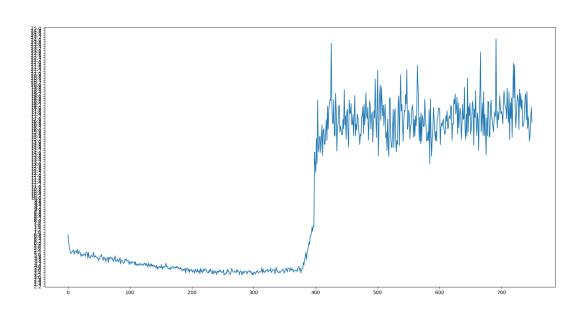


図 4: ASPEC-JC 実験誤差

両方の実験も果たして、よいではないパフォーマンスで表示されているので、来週は誤差を減らすように実験を 進む

4 来週目標

- ManyThings の ENG-CMN データにより実験誤差を減らすこと
- ASPEC-JC データにより実験誤差を減らすこと