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.

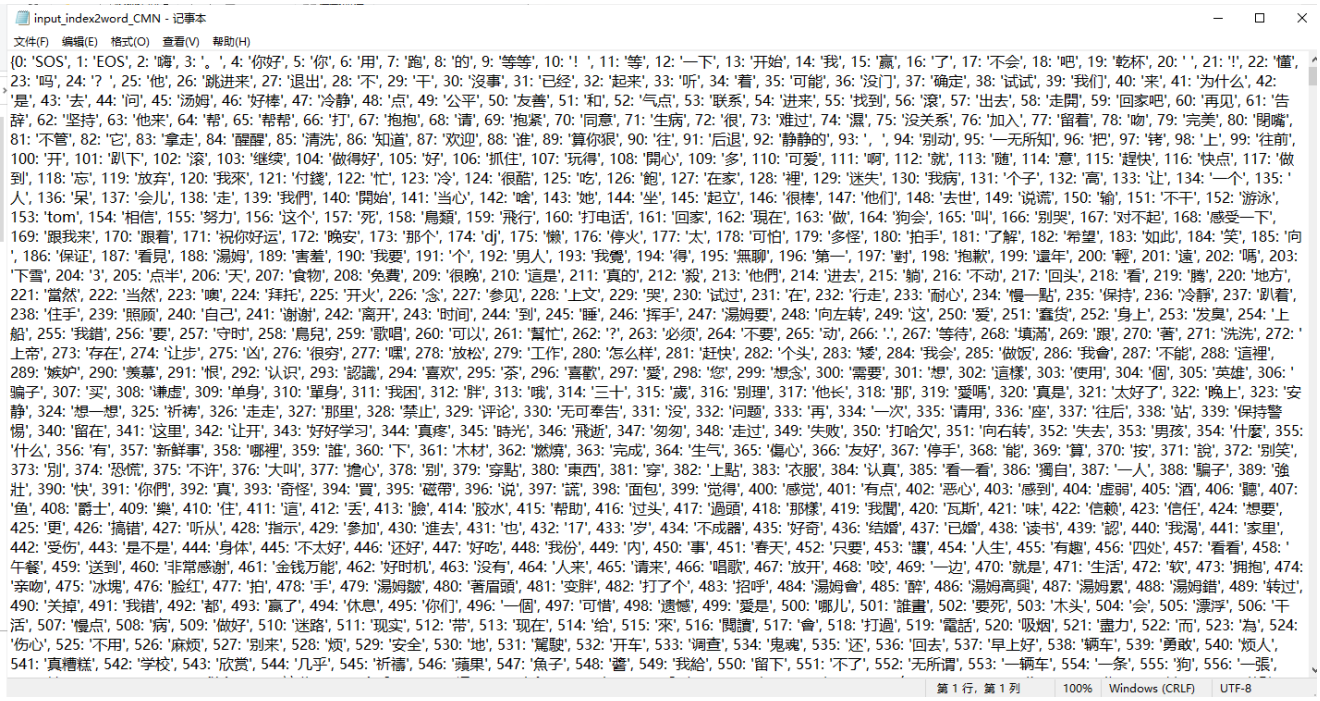

図 1: 中国語 index2words の例

図 2: 日本語 index2words の例

3　実験を行う

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.

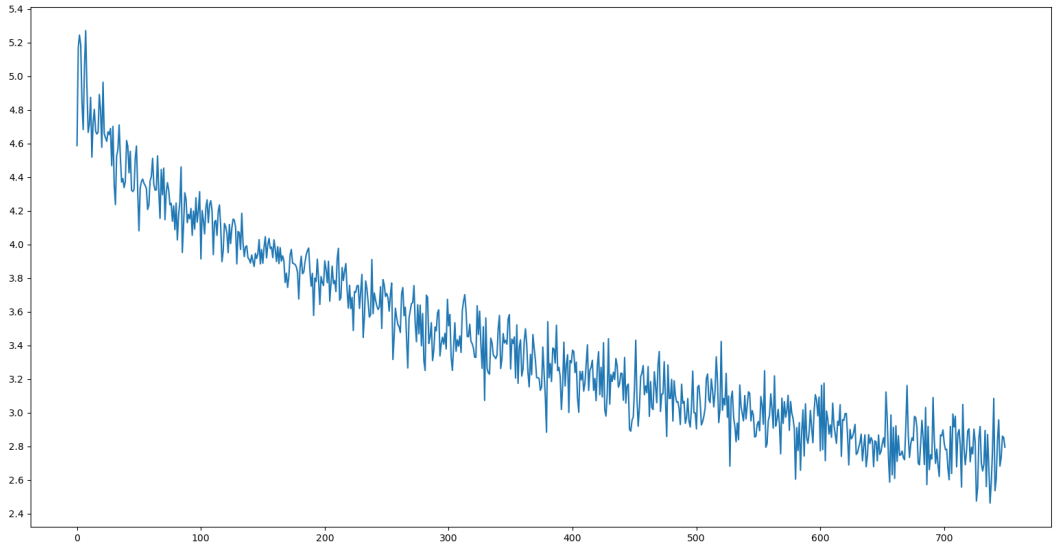

図 3: ManyThings 実験誤差

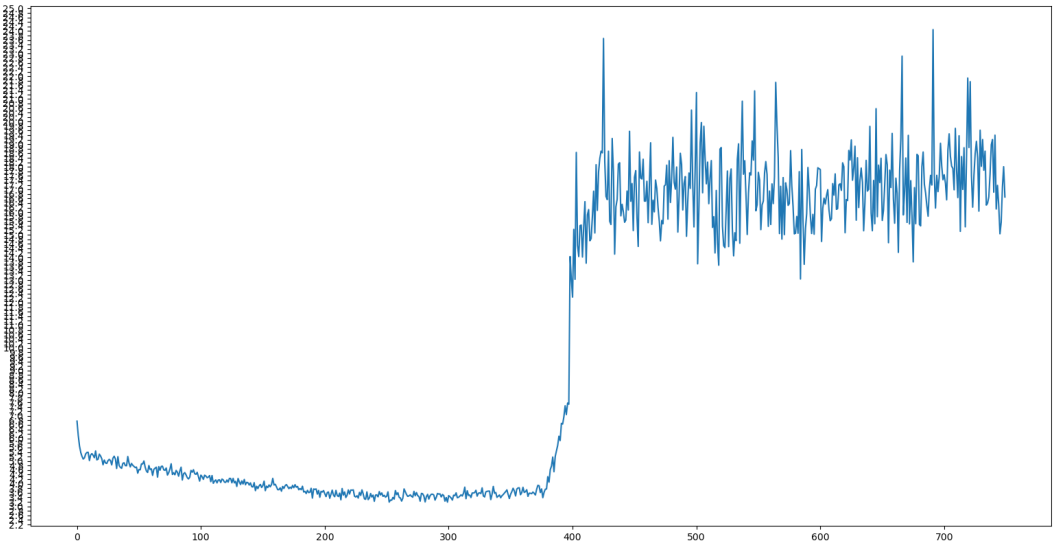

図 4: ASPEC-JC 実験誤差

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