進捗報告

1 今週やったこと

- 実験内容のまとめ

2 要素技術

2.1 seq2seq

seq2seq は [1],2014 年に Google が発表した, 言語モデルである. 従来の DNN(ディープニューラルネットワーク）が扱いにくい時系列データ問題を解決するため,seq2seq は RNN を用いた Encoder-Decoder モデルを導入した.Encoder は入力する時系列データをベクトルに圧縮し, そのベクトルを Decoder に渡し出力系列を生成する.RNN を利用したため,Decoder の出力は自動的に調整する. 始めと終わりのシンボルとして, ＜ SOS ＞と＜ EOS ＞のトークンを導入した, この中に, ＜ SOS ＞は始めの信号になり, ＜ EOS ＞は終わりの信号になる.

2.2 RNN

本実験は,RNN を導入したエンコーダーとデコーダーで行う. 通常のニューラルネットワークでは, あるレイヤの出力は, 次のレイヤの入力として利用されるが,RNN では, 同じニューラルに対して, 当時刻の入力だけでなく, 前時刻の出力も入力し, その出力も次の時刻の入力として扱い, 再帰的構造を持ったニューラルネットワークである.

2.3 LSTM と GRU

RNN は勾配消失と重み衝突の問題が存在するので, 長期的な特徴を学習できない, その問題を解決するため,LSTM と GRU を導入する.LSTM は, 誤差を保存するセル（CEC）と入力, 出力ゲート構造を使用する. 勾配消失問題により, 専用の記憶セルで長期的な記憶を可能にすることで対処する. そして, 重み衝突問題は, 入力レイヤから再帰セルに入力される信号や再帰セルから再帰セルに入力される信号を適切に処理するゲートを用意する.GRU セルは,LSTM と同様の性能を持つ,LSTM より計算量が少なく, 高速に学習を進めるセルである. 本実験は GRU を導入した Encoder-Decoder 構造を使用する.

3 データセットの紹介

ManyThings データセットはネットで収集され, 英語と多言語のデータである. 収集されたデータは英語-他言語のペアを, 単語の少ない方から多いの方までソートされるデータセットである.

ASPEC は,JST と NICT 共同で作るの論文抽出コーパスである, このコーパスは日本語-中国語と日本語-英語二つのデータセットがある. 日本語-中国語データセットは 680,000 日本語と中国語の文章ペアがあり, 日本語-英語データセットは 3,000,000 日本語と英語の文章ペアがある. 本実験は ManyThings の英語-中国語データセットと ASPEC の日本語-中国語データセットで実行する.

4 jieba と janome の紹介

英語とは違う, 日本語と中国語は単語を区別する空白がないので, 実験前に分かち書きの処理しないと, 文章はこのまま seq2seq モデルに入り, 実験はうまくいかないので,jieba と janome でデータを処理する.

jieba は 2013 年にリリースされ, 中国語文章の分かち書きに専用するライブラリである. 本実験は jieba の cut メソッドで中国語の文章を単語に分ける.

janome は Python で記述された, 辞書内包の形態素解析ライブラリである. 本実験は janome の tokenizer メソッドで日本語の文章を単語に分ける.

5 実験の流れ

5.1 データ処理

まずはデータセットの処理. データセットの処理は,jieba（中国語データ）,janome（日本語データ）で行う.

処理したデータは,word2index（単語からインデックス）,index2word（インデックスから単語）,word2count（単語から単語が出現した回数）というディクショナリの形式で保存する. ボキャブラリー数は n_words に保存する. 表１に各データセットの文章ペアを示す. 表２表３に各データセットの言語ごとの単語数を示す.

5.2 モデルの実装

実験に使う seq2seq モデルの Encoder は,RNN である. 入力と前時刻の hidden 状態を入力した時,Encoder は現時刻の出力と hidden 状態を出力する, 出力した現時刻の hidden 状態は, 次の RNN の入力として次の入力と一緒に入力する. 最後の慎吾＜ EOS ＞を入力すると,Encoder の運作は停止し,Decoder の運作が始める.

実験に使う Decoder は RNN の上に,Attention Mechanism を導入したモデルである.Decoder は入力, 先回のヒドゥン状態, エンコーダーの出力を入力とし, 出力, 現時刻の hidden 状態を出力とする.

5.3 トレーニング

トレーニングする前に, データペアを入力 tensor と目標 tensor の形式に変換し, そして入力 tesnsor を seq2seq モデルに入力し, 出力のデータも tensor の形式である.

モデルをトレーニングする時,Encoder 毎回の出力と最後の hidden 状態を追跡する. そして＜ SOS ＞トークンを Decoder の最初の入力として入力し,Encoder 最後の hidden 状態を Decoder 最初の hidden 状態にする. 実験誤差は,seq2seq モデルの出力 tensor と, データセットの目標 tensor により,NLLLoss で計算できる.

6 来週目標

- 誤差を減らす方法を探す
- testing data を導入する

参考文献

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, Vol. abs/1409.3215, , 2014.

表 1: Pairs

DataSet	ManyThings	ASPEC-JP
Pairs	24360	2107

表 2: ManyThings_Words

ManyThings	ENG	CMN
Words	7484	14716

表 3: ASPEC-JP_Words

ASPEC-JP	JPN	CMN
Words	6640	6844