夏季レポート 4

情報科学類 3 年 江畑 拓哉 (201611350)

Contents

1 CopyNet の理解とその応用についての考察

より柔軟に対話システムを作ることが出来るらしい CopyNet とそれに関連して意図的にデータのノイズを乗せる手法について調べた。

1.1 CopyNet

1.1.1 概要

Seq2Seq に対して、入出力時のコピーを行うタスクを追加したもの。例えば、任意の A に対して「A はどうですか？」という質問に対して「A はーです。」といった回答を期待することが出来る。

つまり任意の対象について (例えばある任意の人物の評価について) 議論する場合にこの手法は有効であると考えられる。

1.1.2 実装

PyTorch での 実装 があるため、これを熟読している。

1.1.3 問題点

任意のそれに対して有効なので、特徴をもたせる際に問題となる可能性がある。

例えば、A は好きで B は嫌い、という特徴を持たせたいが、「X は好きですか？」という質問に対して「X は好き/嫌いです」という回答を生成してしまう場合、どちらかは間違った回答を生成してしまう。

1.1.4 注目した理由

CopyNet は対話タスクや要約タスクにおいて単語をコピーするという手法を取っている。また、同様の手法として Joint Copying and Restricted Generation for Paraphrase に示されているものがあり、これに関してはまるで単語をコピーしているため、未知語に対しても同様の結果を得ることが出来ている。

しかし日本語での実験結果は この ようなものであり、対話というにはやや心もとないと考えられる。

ところがスタイル変換というタスクでこれを見た場合、この手法は極めて有力な手法のように思える。

特に今回求めているスタイル変換ではかなりの確率で Copy を要しているため、これを使うことで良い結果を得られるのかも知れないと考えている。

1. 実際の例 （ここ より引用）

(入力)	(出力)
おはよう	おはようございます
調子はどうですか？	調子は
お腹が空きました	お腹が空きました
今日は暑いです	暑いです

1.1.5 参考文献

Incorporating Copying Mechanism in Sequence-to-Sequence Learning ALC2016 読み会@すずかけ台
Incorporating Copying Mechanism in Sequence-to-Sequence Learning
Joint Copying and Restricted Generation for Paraphrase

1.2 Unsupervised Machine Translation Using Monolingual Corpora Only

1.2.1 この論文に至った経緯

先述の CopyNet における話で、“コピーによって未知語に対する反応が出来ている”という部分、そしてどうやらデータセットをそこまで大きくすることが時間的に難しいという問題、そしてスタイル変換に関する別の実装を示した、 Style Transfer from Non-Parallel Text by Cross-Alignment の実装を眺めていた所偶然“データにノイズを乗せる”という手法を発見したことによる。

1.2.2 この論文の概要

（※まだ精読が出来ていないため不安定な部分が含まれている）
機械翻訳のタスクにおいて並列なデータを要している問題に対して、中間表現を作成することで非並列なデータでこのタスクを達成できるようにしようという取り組み。中間表現を作成するために AutoEncoder の中の、特に学習時に入力に対して任意にノイズを乗せる Denoising Auto-Encoding を用いている。

1.2.3 DAE とは何か

DAE こと Denoising Auto-Encoding は AutoEncoder の一種で主にデータにノイズを乗せ元のデータを復元することで逆に元データのノイズを削ることが出来る手法であり、日本では専ら音声認識のノイズ除去に使われているようである。
詳しい説明を Stacked Denoising Autoencoder の汎化性能向上に関する一検討 から以下に引用する。

`Denoising Autoencoder(dA)` は、上述した `Autoencoder` の学習段階で、入力データにノイズを付与し、そのノイズが付与する前のオリジナルデータを復元するように学習を行う `Autoencoder` である。このようにして、 `dA` は、単純な `Autoencoder` と比較して、より頑健な特徴抽出が可能となり、未知のデータに対する汎化性能を高めることが出来る。

2 Twitter を用いたデータ収集について

Twitter よりデータを収集する手法に関する記事が最近公開されていたため、それについて学習している。

2.1 データ収集について

API を叩くことでデータを収集することが出来るようである。
この部分に関しては既存のものがあるためそれを利用する予定である。

2.2 前処理について

この 前処理手法に加え、手でデータを精査、再精製する必要があると考えられるが、（現実問題として自分の脳からデータを作成するのも（発想力の問題で）限界に近いため）こちらのほうがデータの生成速度を挙げることが出来ると考えている。

2.3 参考文献

Twitter から対話データを収集する

3 日本語の類語について

3.1 WordNet について

日本語を取り扱う上で厄介な問題として同義語や、漢字・かな問題がある。漢字・かな問題に関してはすべてをひらがなにすることで対応が出来るが、同義語に関してはどうだろうかと調べ、WordNet にたどり着いた。

WordNet とは語の類義関係のセットをグループ化し、各グループの相互関係を結んでいるネットワークを検索するツールである。これを用いることで同義語を一つにまとめることが出来るのではないかと考えていた。

3.2 実際に WordNet を使ってみた結果

ご飯	供米, 舎利, 白米, 神米, 米, 八木, 禾穀, ご飯, 米穀, 稲孫, 飯米, 米飯, 上げ米, 枕米, 藪稲, 御飯, おまんま, 糧米, 上米, 田の実, 飯, 粮米, 銀飯, 穀, ライス, 産米, 稲, イネ
おはよう	なし
こんにちは	やあ, こんにちは

以上の結果から、ある一定の精度で良い結果を得られることがわかった。
これを用いることで大幅な語彙削減を行うことが出来る可能性があることがわかった。
問題点として、細かいニュアンスの違いを表現できない可能性があるということである。例えば、“ご飯” を “食事” というニュアンスで用いている場合、“イネ” は余りにも不適当な表現であるはずだ。
また、どの単語をメインの単語にするかは難しい問題だろう。

4 今週の方針

- 手打ちデータを 1000 まで増やす
  現在約 800 まで集まったので更に増やす

- Twitter からデータを収集する
  実際にデータを 1 万ほど集めてみる

- CopyNet を書いてみる
  実装したものが既にあるため、これを書き写すことになると思われる

- Style Transfer の実装にノイズ部を書き足してみる