進捗報告

江畑 拓哉 [1]

2018-12-12 Wed

[1] s1611350@u.tsukuba.ac.jp

全体像

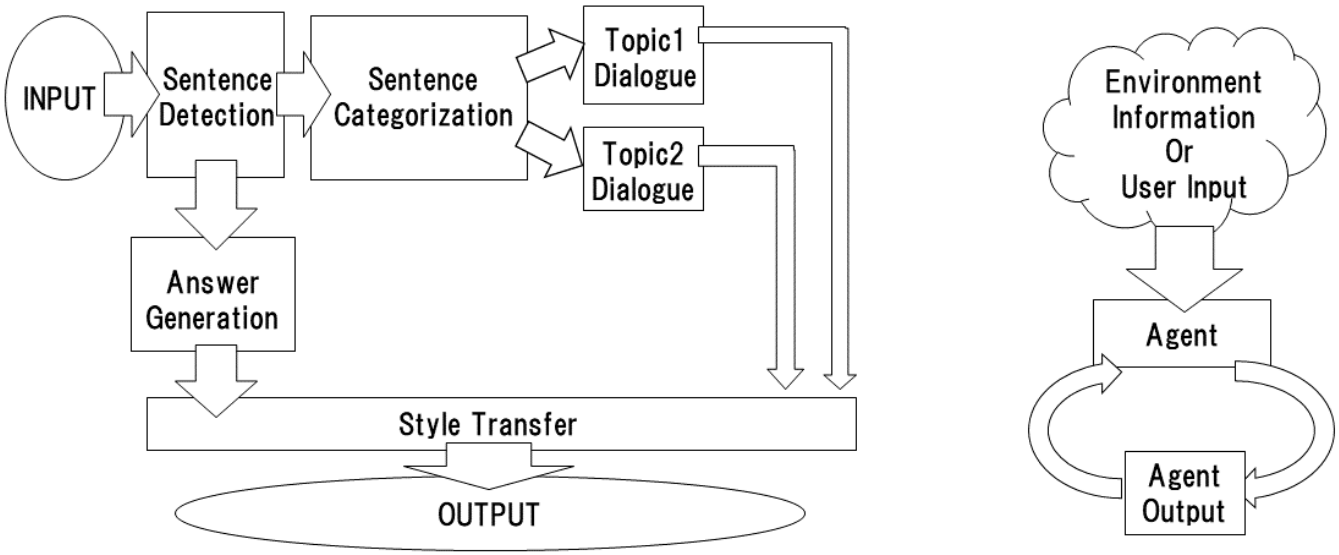


Figure: 全体像

研究背景

- 主目的： 前スライドのシステムに基づいた対話システムを作成する
- 既存研究: amazon alexa prize
  毎年行われる amazon alexa の内部システムを競うコンテスト。昨年度も今年度も 全体像 のようないくつかのモジュールに分割したシステムを構築したシステムが優秀な成績を収めている。
- 取り組む問題:
  - 各モジュールを一つの問題として捉え、それぞれを解く。
  - 各モジュール間の出力を繋ぐための問題を解く。

今回報告する部分

⋆ (): 対応するモジュール

- スタイル変換 (Style Transfer)
- 不均衡データ分散なデータセットと不均衡データ数なデータセットの関係 (Sentence Detection)
- 日本語言語理解タスクのための Subword
- 今後の実験計画

スタイル変換

- センテンスにおける口調を変換する という問題。
- 今回主に取り扱うのは、 です・ます調<=>話し言葉 への変換。
- データセットは自作の 並行な ものを使用 (データ数が少ないので Loss や Acc への言及はしない)
- 3 つの手法で実験し、その結果を観察する。

Sequence to better sequence

- 対象とするデータ: 非並行な 2 つのスタイルを持つセンテンスの集合

学習内容

1. 入力文と出力文を一つの VAE に通し潜在表現を得る
2. 入力文のスコアを 0 、出力文のスコアを 1 となるように潜在表現を評価する FF を学習する
3. 何らかの入力 $\boldsymbol{x}$ に対してそのスコアを 1 になるように潜在表現を調整することで、スタイルが付与された文 $\boldsymbol{x'}$ を得る。

実験ではこれと、更に多くの汎化性能を得るため、VAE の入力にノイズ (一部を未知語に変換する) をかけたものの実験を行った。

CopyNet

- 対象とするデータ: 並行な 2 つのスタイルを持つセンテンスの集合

学習内容

1. 入力文を通す RNN から最後の出力 (意味ベクトル)・それぞれの単語を表すベクトルを得る
2. 出力文を生成する RNN と演算を行う
3. 出力文を生成できるようにそれぞれのベクトルに対する重み付けなどを学習する

本来は 機械翻訳 のタスクで使用されている手法だが、今回対象とする問題の、 入力文・出力文が非常に似ている という性質からこの手法を用いた。

実験結果

実験結果

実装	入力	出力
S2BS	おはようございます。	おはよう。
	応援する。	応援してる。
	今日は寒かった。	今日は寒かった。
	夕飯は？	夕飯はどうしようか？
	早く寝たい。	早く寝た方がよいね。
S2BS with DAE	S2BS と同じ	
CopyNet	おはようございます。	おはよう。
	今日は良い天気ですね。	今日は良い天気。
	こんにちは。	こんにちは。
	頑張るぞい！	頑張るぞい！
	進捗どうですか？	進捗どう？

考察

- S2BS と S2BS with DAE にはほとんど差異が見られなかった。

これが データ数 によるものなのか、本当になんの意味もないのかは不明である。

- S2BS、CopyNet のいずれでも 入力が正規化されていなくても 問題なく変換ができることがわかった。
- S2BS は CopyNet に比べて 表現力が大きい ように感じられる。
  → 学習内容から用意に予測できる。

不均衡データ分散なデータセットと不均衡データ数なデータセットの関係

本来の問題: 任意の入力文からいくつかの質問・文を抽出したい。

しかし以下の理由から画像認識の問題として (1) の問題を設定した。

- 自然言語 (日本語) の データを十分 に用意できない。
- 2 つの選択肢が考えられる。
  1. シンプルなクラス分類 (文 A、文 B . . . その他)
  2. 文章類似度 を活用したクラス分類

$$
\begin{aligned}
class &= \max_i f(similarity(x, Y_{ij})) \\
x &\ldots input\ sentence \\
Y_i &\ldots set\ of\ sentences\ in\ class\ i
\end{aligned}
$$

設定した問題

ネコ画像の集合 X と、イヌの画像の集合 Y 、ランダムな画像の集合 Z を用いる。
$X-Y$ $X-Z$ の 2 値分類問題において、データ数の比率を変化させながらその Loss, Acc を比較する。

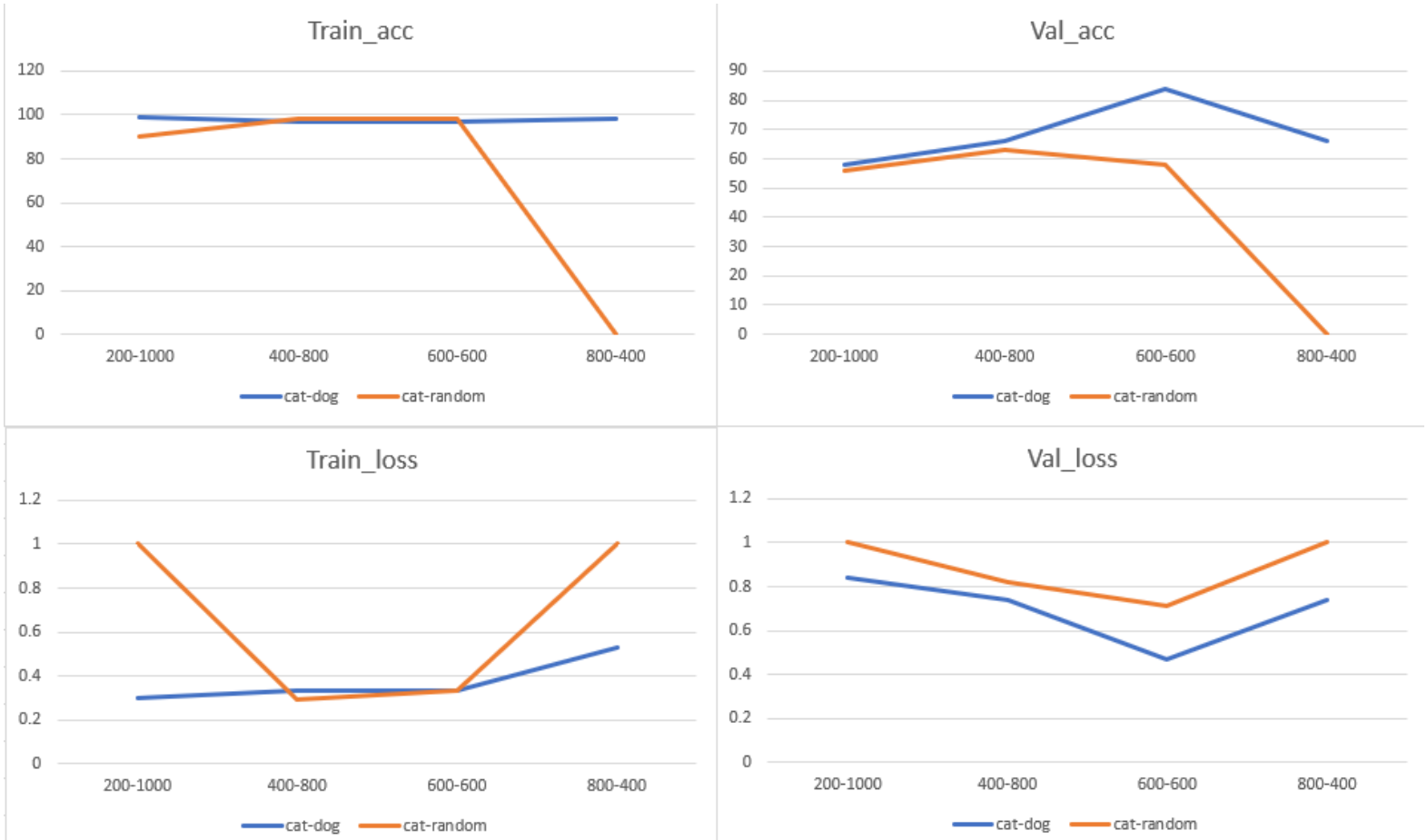
実験

6 層 CNN を用いて実験を行った

- 入力画像は 28×28 の 3 チャンネル
- 出力は $2 \times batchsize$ (多クラス分類への拡張を想定しているため)
- データ数の比率 $x : y$ に対して $y : x$ となるように loss の重み付けを行った。
- 最適化関数は Adam を用いた
- epoch は十分に学習ができるまでとした
- 検証データのデータ比率は 1:1 とした。

実験結果

以下の通り (loss = 1, val=0 は値が荒れて計測が出来なかったことを示している。)



考察

- 全体的に ランダム画像 とのクラス分類の方が 精度が悪い とわかる。
- ランダム画像は与えられればその分精度が上がると考えていたが、そのようなことはなかった。
  - ランダム画像を増やした場合の精度の変化の一部は、イヌ画像の精度の変化に近いものがある ことがわかった。
- loss の重み付けをより極端にして実験を行ったが、結果はほとんど変わらなかった。
- 1 : 5 のデータ比での実験は殆ど結果を得ることが出来なかった。(学習が出来なかった。)

日本語言語理解タスクのための Subword

- Subword とは単語分割を行う手法の一つであり、一般的な単語分割より語彙数を減らすことが出来る。
  例えば “subword → sub + word” を挙げることが出来る。
  一般に機械翻訳の分野で用いられているが、日本語の場合では上手く分割することが難しい。(ほとんど単語分割に近くなってしまう。)
  更に英語などで用いられているプログラムでは漢字かな入り混じり文のせいで上手く分割できない。

問題設定

カテゴリ分類やその他の機械学習を用いた自然言語に関する問題を解く際に どのように文を分割すれば良いのだろうか。

- 機械翻訳では subword は優秀だが、言語理解などではあまり優秀ではない。
- 漢字とかなが入り混じっていることで日本語の学習は無駄に難しくなってしまっているのではないだろうか。
  → かなのみの方が subword という意味では計算しやすい。

実験

fasttext の skipgram を用いて単語の分散表現を得る問題で、漢字かな入り混じり文、かなのみの文の 2 種類を用いて性能比較を行う。

- データセットは十分なデータを用意するために wikipedia のログを用いた
- fasttext の実装は公式が発表しているものを用いた
- loss はそれぞれの単語の分散が正しいのか (つまり意味的に近い単語が近い位置にあるのか) を計算して求めている。
- 計算される類似単語を比較した

実験結果

比較として、“日本 (ニホン)” を用いた

漢字かな入り混じり文

韓国
米国
台湾
にっぽん
中国
日本さくらの会
海外
実業
国内
日本税理士会連合会

かなのみの文

ニホンヤモリ
ニホンバレ
ニホンシカ
ニホンウンソウ
ニッポンザル
ニホンズイセン
ヒトツオボエ
ゴジセイ
ニホンカジョシュッパン
ニホンドケン

実験結果

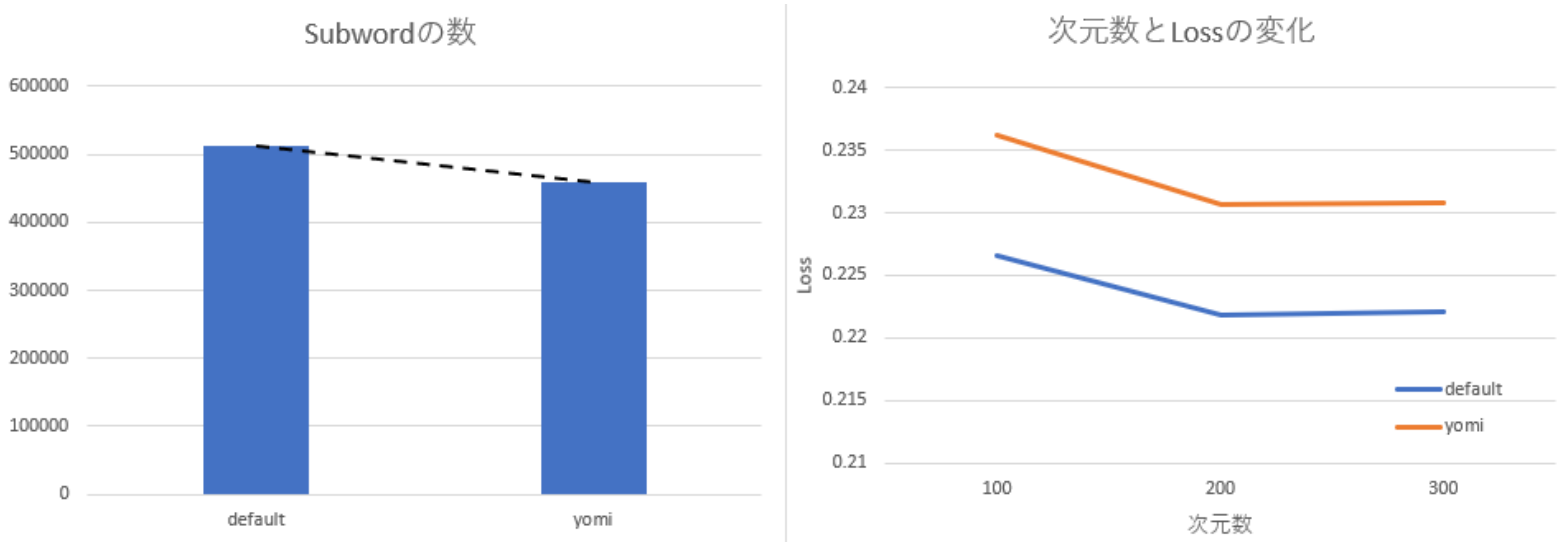


Figure: subword と漢字・かな/かなの関係

考察

漢字かな入り混じり文は国として類似する単語を取り出していることがわかるのに対して、かなのみの文では 生物名 や日本晴れ、といった 慣用的な表現 を多く抽出している。どちらが良いのかを決めることは難しいと考えられる。ただカテゴリ分類の立場に立つのであれば、おそらく前者のほうがより良い結果を導けるのではないかと考えられる。

今後の実験計画

- 極性判定

日本語言語理解タスクのための適切なフォーマット の項で議論できなかった極性判定について同様の実験を行いたい。

- CoLA タスクを用いた自然言語判定
  ある文 が自然なものであるかを判定する CoLA タスクを解く問題を あるモデルから出力される文 に対して同様に処理できるのかを調べる。これは各モジュールを接続するための問題として掲げている。
- 文章類似度を用いたクラス分類
  不均衡データ分散なデータセットと不均衡データ数なデータセットの関係 の項で紹介した (2) の文章類似度を用いたクラス分類について調べる。
- 1:1 対話
  入力文 1 文に対して、前後の文脈は考慮せずに出力文 1 文を