○○風の文書への
スタイル変換

2018/3/1
B3 プロジェクト研究B

伊藤智彦

1.目的

標準語

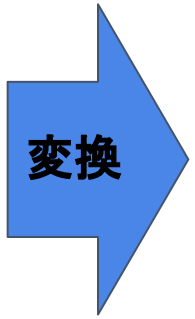
「私は猫です」

標準語

「あなたに会いたいです」

標準語

「あなたは馬鹿ですか？」



”夏目漱石”風の文章

「吾輩は猫である」

”メンヘラ”風の文章

「会いたくて震える」

”アスカ”風の文章

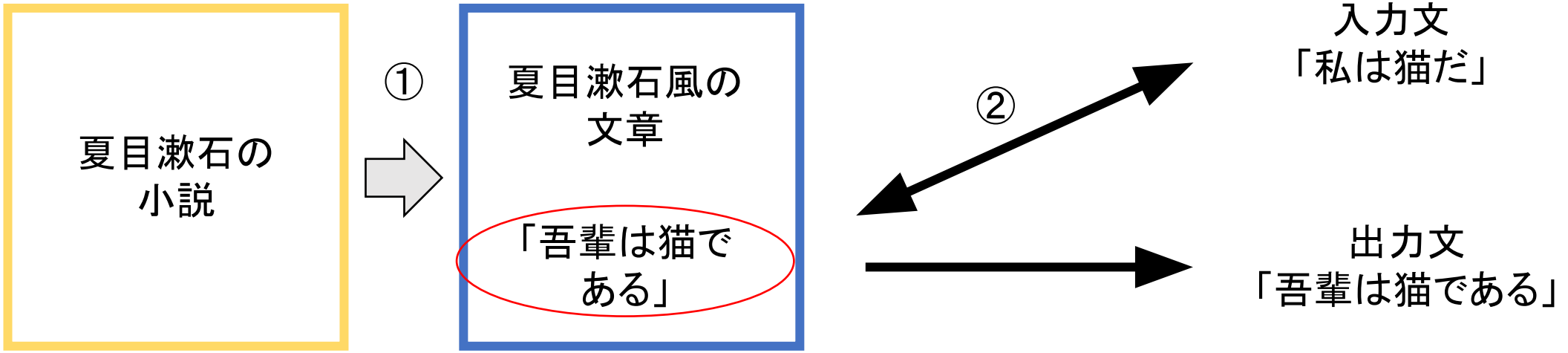
「あんたバカァ！？」

2.実現方法 (自動生成+類似度計算)

学習データから

①学習データ風のスタイルを持った文章を自動生成し，

②自動生成した文章の中から，入力と意味的な類似度の近いものを探す

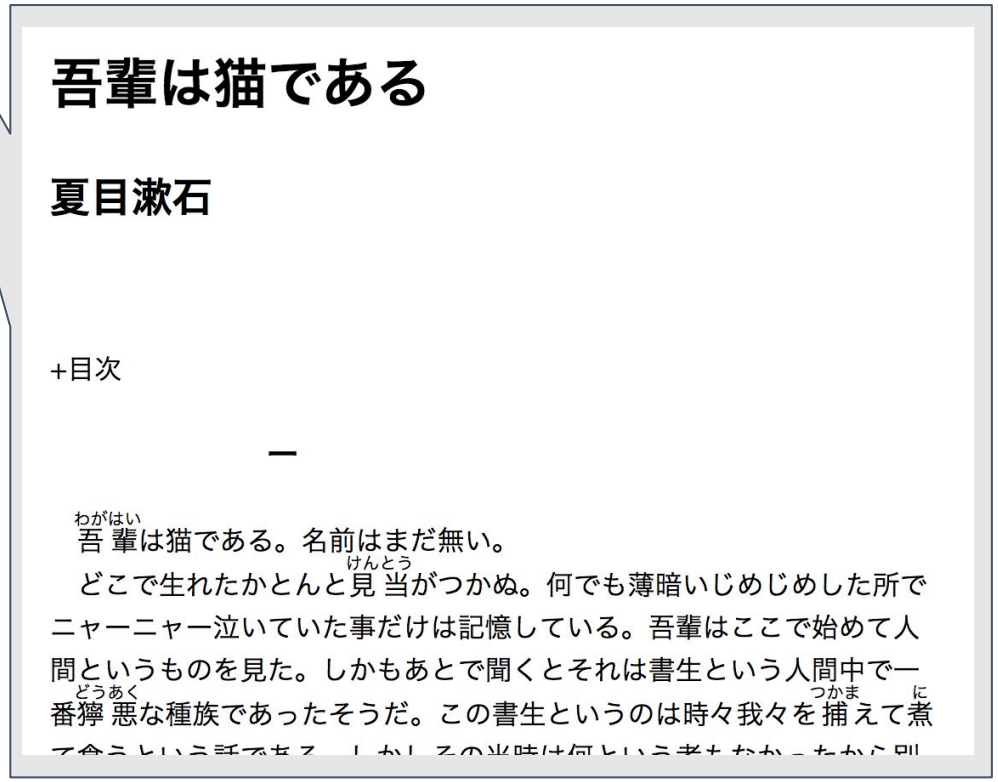


3.1 実験その1 学習データ

wagahaiwa_nekodearu.txt（「吾輩は猫である」の全文）

青空文庫に夏目漱石著「吾輩は猫である」の全文が載っている！ありがたい

http://www.aozora.gr.jp/index_pages/person148.html



吾輩は猫である

夏目漱石

+目次

一

吾輩（わがはい）は猫である。名前はまだ無い。

どこで生れたかとんと見当（けんとう）がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。吾輩はここで始めて人間というものを見た。しかもあとで聞くとそれは書生という人間中で一番獰悪（どうあく）な種族であったそうだ。この書生というのは時々我々を捕（つかま）えて煮（に）

3.1 実験その1 自動生成の結果

繰り返し回数: 1 - loss: 4.1994

申しょいを少しも充に落ちる語ったかど勢だ」「いらせんと従ょいに詩に至った文考を逢ってかかり方でいし出である武び少しドャ大碗屋に見て、這入ったが内にステ動りら間」
「人のく迷亭はかあした共安である。あるむを御当坊の起粧にな通ったの云き返す小ッちにも通ろ首の朝な性に学者口を高「え最がつけない」「すん、毫は食うんだ流山主人はレーの子口学者は奥ずや。あ昔威横あが、しいくオあした　活で、学種爺とにこの作団だからパ気と云いえ面を見た。

繰り返し回数: 27 - loss: 3.3512

大きな声を第一のはなかったから、それだから、どうかした事がない。その時の方だからしきりになっている事はないが、これからその事だから、自分でも、そのそとでない。吾輩は人間になると、吾輩には文明のごとくのはない。一るのは一際もなくなる。それなら飛び出して来たものがて来たのだが、またはなかったが、それではあいまさに、云わぬが、なるほどのところがある。ただ一人がいい。

3.2 実験その2 学習データ

まず”**メンヘラ**”風の文章とは…

”メンヘラ”の定義を調べると

- 寂しがり屋
- ヒステリック
- ネガティブ
- 感情的になりやすい　etc…

参考：生活百科, メンヘラ女の14個の特徴と注意点,(https://seikatsu-hyakka.com/archives/2511)

Love Recipe, メンヘラの女性に共通する特徴や恋愛事情,(https://love-mag.jp/columns/524)

3.2 実験その2 学習データ

まず”**メンヘラ**”風の文章とは…

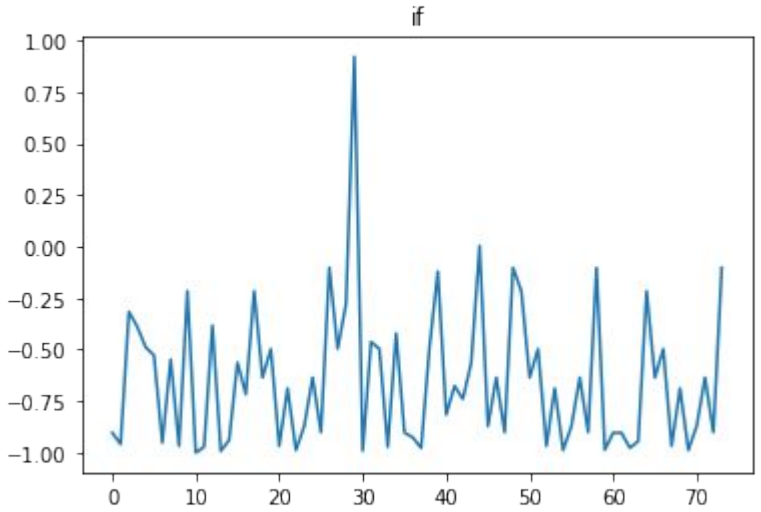
先ほどの”メンヘラ”の定義を踏まえて考えると

西野カナ

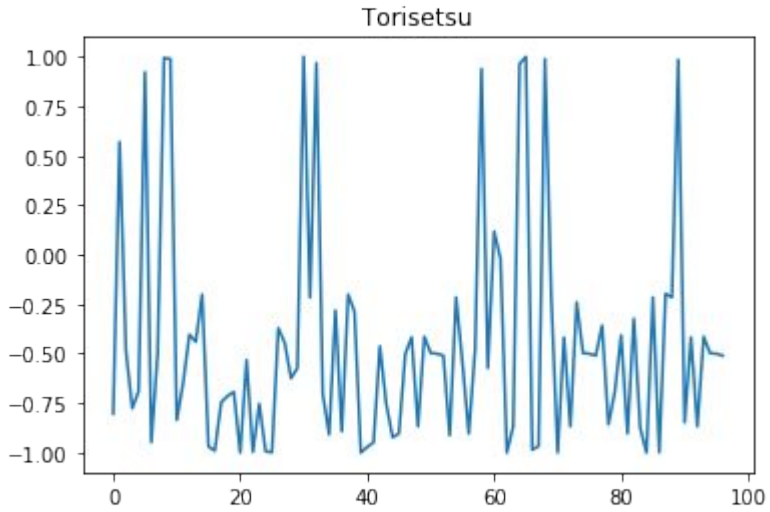
の歌詞に着目

3.2 実験その2 学習データ

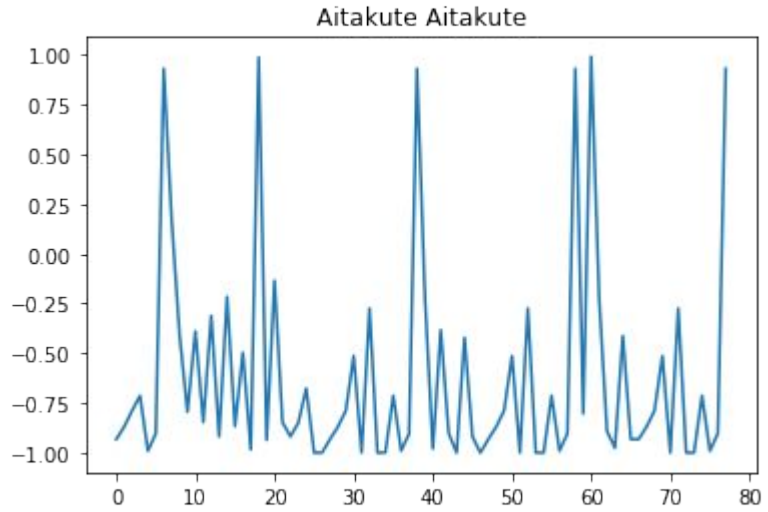
西野カナの全歌詞データ(kana.txt)をスクレイピングで集め、
集めたデータから曲ごとに感情分析を行うと…



if



トリセツ



会いたくて会いたくて

スクレイピングをしたサイト:歌詞検索　J-Lyric.net (http://j-lyric.net/)

3.2 実験その2 自動生成の結果

今 は あと もう少し の 思い出 の 日 も 笑え ない のに そういや 私 も その 笑顔 その 瞳 の こと 君 と 君 た ずっと 君 が 笑え て くれ た よ どんな 時 だっ た だっど 私 は い たい ね とも 何 誰 もも もと 君 無け で 見せ て いる こと は 忘れ ない だ らん も人 君 が いい よ そう 君 が いな でしょ あなた は こって ね 私 が ある から 朝 からし 強 ない あの 子 な の こい 心 見つさ にる のに 一緒 に もし て 同じ 景色 を できし たい	どんな 悲し も 君 の いよ この 気 は どう けど あっ た 大人 だ よ ね み て なら 大人 だ ん だ から 優しい でこょ れ で いついける から もんないん だ から も かんな も カ で は もっと ちまっ て い たら 一人 だ に ね 君 の 二人 の に も ある けら ない から 今 は この 時 も あの 子 に ない た シャし 何イ を 私 の の は 一つ で も リーる なんて ない ままで 何 の の こと を あえ ない 大好き だっ た ん だ
繰り返し回数：59回 diversity：0.2 loss：0.2865	繰り返し回数：57回 diversity：0.2 loss：0.2909

3.2 実験その2 考察

○精度が低い理由

・歌詞データだと前後の関連は高いが、もともと文法的におかしいところがある。

（ex 繰り返しで同じ単語がある、句読点がないので文の終わりが曖昧）

・学習データが少ない

・英語と日本語が混じった文章で英語だけを取り除くと、残った日本語の文章がおかしくなり形態素解析がうまくできないことが多かった。

3.3 実験その3 学習データ

・アスカ.txt (アスカの台詞)

アニメ「新世紀エヴェンゲリオン」全26話のセリフからアスカの台詞のみを抽出した。

http://lovegundam.dtiblog.com/blog-category-7.html

・アスカ訳.txt

アスカ.txt内の「あんたバカぁ」「あんた」「バカぁ」を

「あなたは馬鹿ですか」「あなた」「馬鹿」に置換したもの

自動生成　：　アスカ.txt

類似度算出:　アスカ.txt, アスカ訳.txt

3.3 自動生成①

以下の辞書作成にあたって，文頭に ”@”を配置した

dic[][] = {単語:出現確率}

マルコフ連鎖

学習データ

@今日は学校に行く日だが

台風なので

私は学校を休む。

@学校は好きだ。

dic[“@”] ={“今日”:0.5, “学校”:0.5}

dic[“@”][“今日”]={“は”:1.0}

dic[“今日”][“は”]={“学校”:0.5, “好きだ”:0.5}

dic[“は”][“学校”]={“に”: $0.\dot{3}$, “を”: $0.\dot{3}$, “は”:$0.\dot{3}$}

dic[“学校”][“を”]={“休む”:1.0}

出力例:「今日は学校を休む」

3.3自動生成①結果

HMMもやればよかった…

レディーファーストよ！
よりにも。
ほんとは加持さんも教えてくればいいのに。
シンジは知りもしなくてはい！
あのバカ！
無への回帰を。
おかえりー。
いやぁぁぁ！
それは碇司令、本当に忙しかっただけよ！
軸線に乗ったわね…退屈なんですもの、ニコニコ笑っていたいわよ！
どいてって言ったでしょう！
ちぇっ！
ミスは許されないじゃ、習ってなかったの？いらないのに！
でも付き合い悪いのよ!でも、今日からお払い箱よミサトは私と暮らすの。
みんな嫌な事もういらないの！

マルコフ連鎖による生成
(確率random)

A.T.フィールドは中和してたのね。
普段からボケボケっとしてる。
冷却液の圧力をすべて三番にまわして！
もう泣かないの。
分かんないのォ！
でもこんなださいの着て、生きてるのね。
優等生も、ラーメンなら付き合うって言うのよね …退屈なんだ、と思い込む！
退屈だからセカンドインパクト世代って、他人に幸せ押し付けないでよ
ほら、お望みの姿になったわ。
ねぇねぇ、加持さんは？ええっ …もう、どうしようもないわ！
分かんない敵を相手に、帰ってきたの！
この3バカトリオが！
分かんない敵を相手に、帰ってきてないわよ。
何が始まったの？ほんとにバカね。
なんで兵器に心なんか要るのよやるわ !変なこと考えないでぇ！

マルコフ連鎖による生成
(出現頻度に応じた確率で生成)

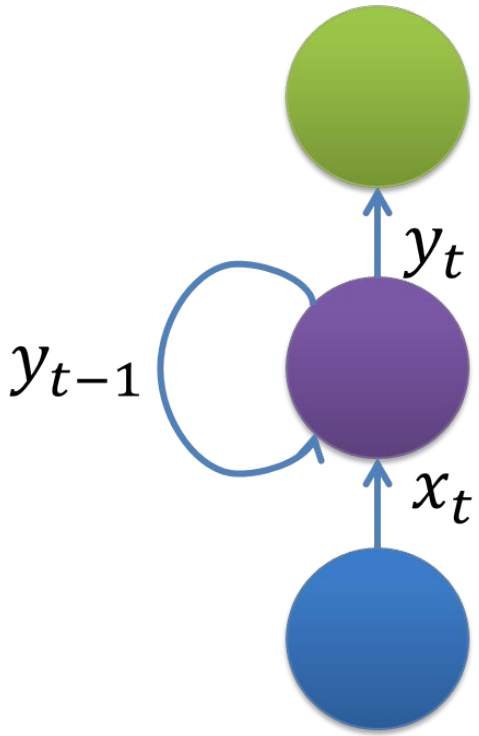
3.3自動生成②

RNNを使うと嬉しいとき

例：

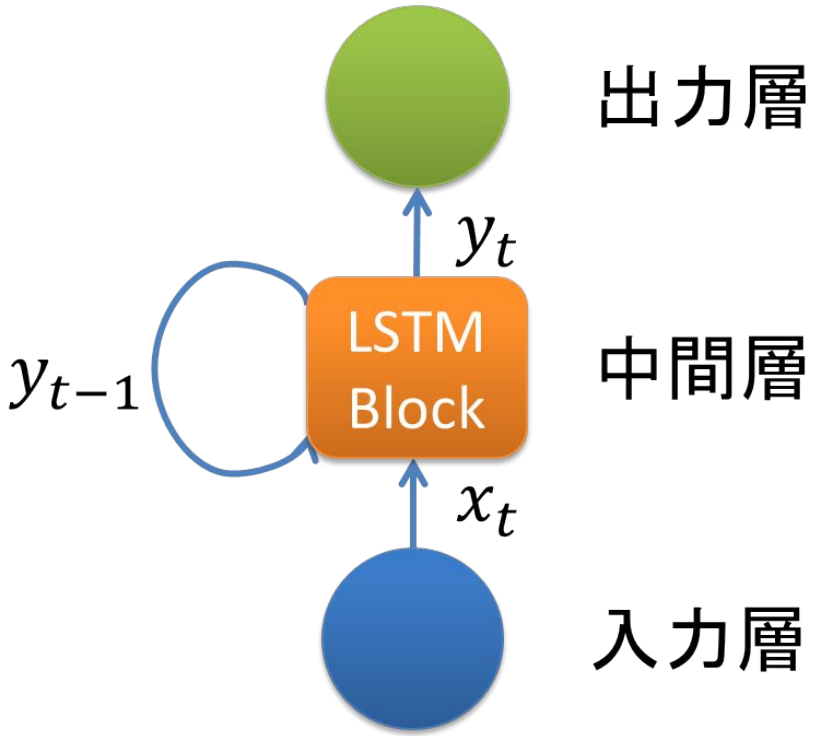
t-n「彼は昨日夜更かししたらしい」

t「 彼は 眠い。」



通常のRNN

→「腹ペコ」を推測するために，nステップ前の情報が必要(再帰により実現)



Long-Short Term Memory

LSTMを使うと嬉しいとき

例：

t-m「私は猫だ。」

t-n「彼は昨日夜更かししたらしい」

t「彼は 眠い 。」

→「眠い」を推定するとき，「私」はどうでもいい。t-m,t-nどちらに大きい重みを与えるかを考慮

引用元:「**わかるLSTM ～ 最近の動向と共に**」https://qiita.com/t_Signull/items/21b82be280b46f467d1b

3.3自動生成②結果

繰り返し回数: 47 - loss: 0.0596

は？
えぇぇ！
きゃがいっ。
あんたもんな、鼻えてそう。
何よ、すまってるいみんぁ、やのTどうしよう一、ぐうくない
あんたバカ？
生きっとのよ!
あんたバカ、エックばっかんなのだからないでしょ！
うるさいわねぇ!シンジは。
そうやう。
私が匂、何き、してくれるすよ。
あんた嫌、…分っ行くせいいやのよ！
へ体バカぁ？見といがんかっただシンジ！

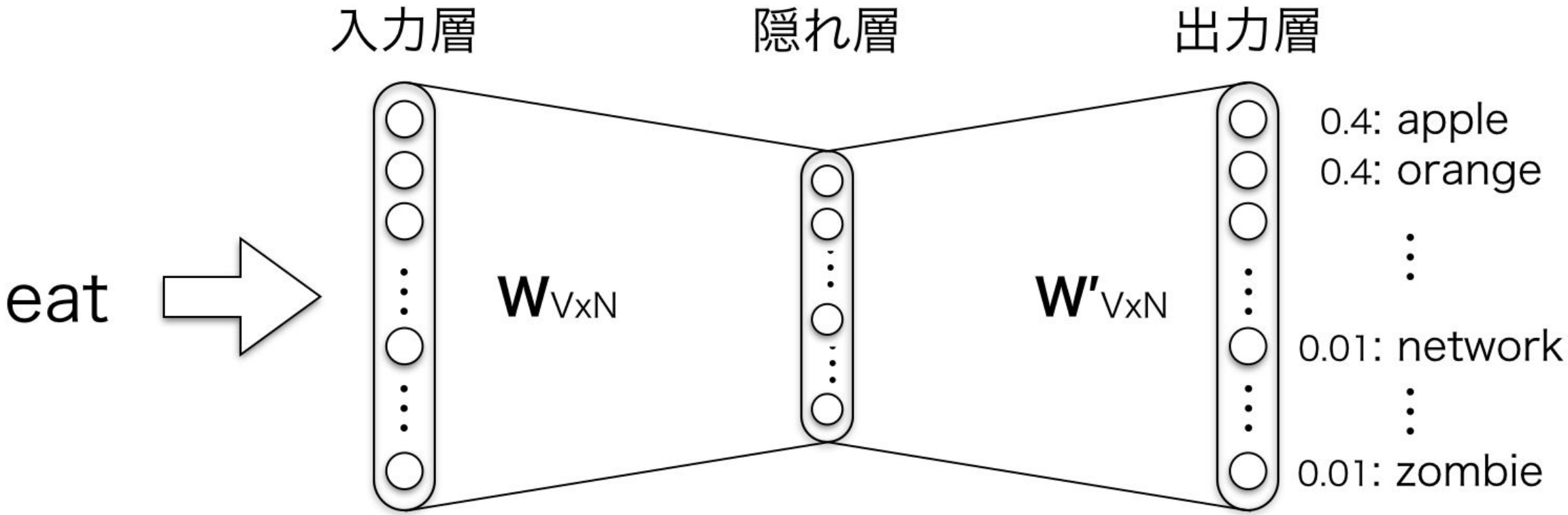
LSTMによる文字レベルの生成

繰り返し回数: 28 - loss: 0.0060

失礼うっ！
なっのよ、シンジはシンジのよ!
あんたバカぁ？あんた、見てるわよ。
正面、私の弐号機にそのするのかしら！
それとも、私のよ！
最後！あたしあたしあたし、こんなかけしなんてない..
何の、できたんじゃない。
だからのよよ。
だから、ラーメンって訳!
あんたバカぁよ!
さあわよ。
あんたバカ～んでしょ、ちゃんとねぇ！
あんた、何のよよ。

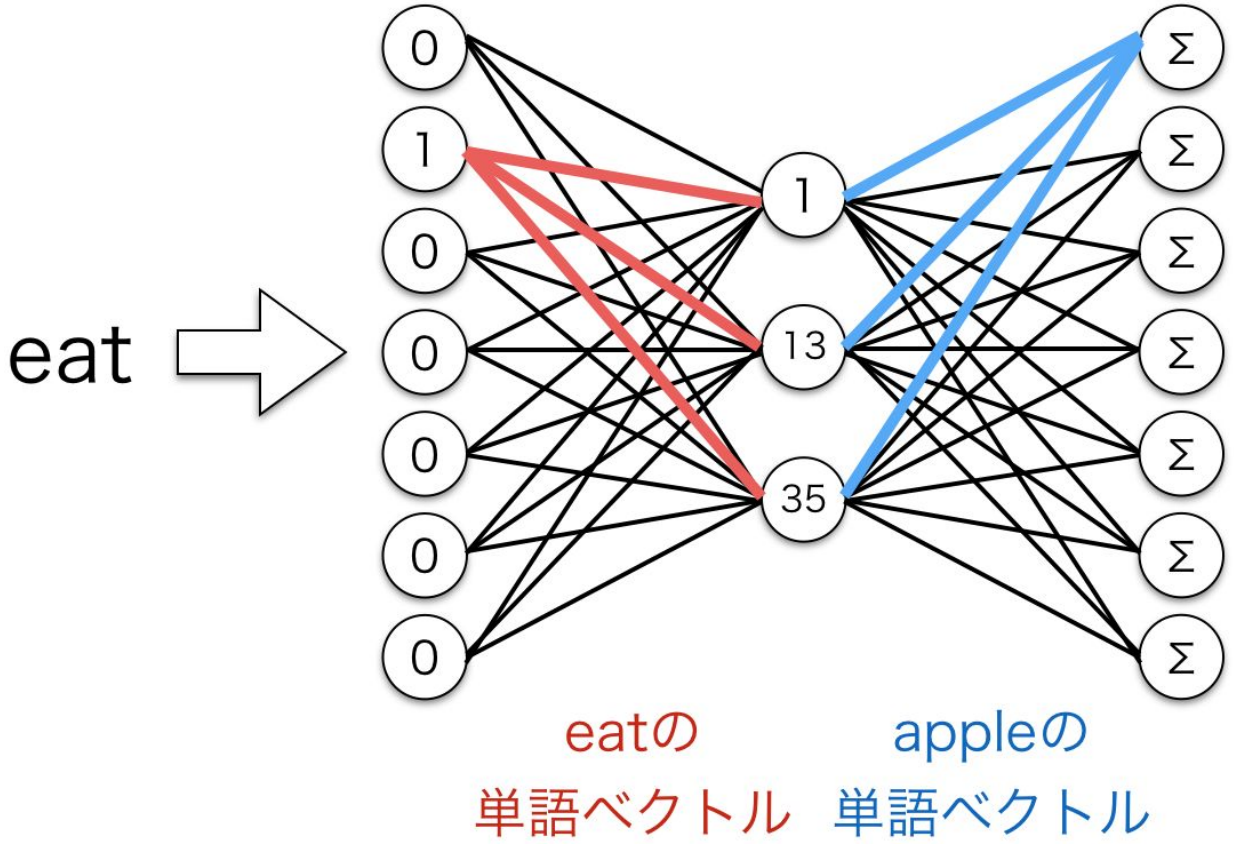
LSTMによる単語レベルの生成

4. 類似度 (Word2Vec)



引用元:「絵で理解するWord2vecの仕組み」https://qiita.com/Hironsan/items/11b388575a058dc8a46a

4. 類似度 (Word2Vec)



0.4: apple

確率値を,
単語ベクトル同士の内積を
Softmax関数$\left(\frac{e^x}{\sum e^x}\right)$でまるめたもの
としている

引用元:「**絵で理解するWord2vecの仕組み**」**https://qiita.com/Hironsan/items/11b388575a058dc8a46a**

4. 類似度算出(単語)

ざっくりとした説明:

似た意味の単語なら周辺の単語分布も似るはず

吾輩は猫である。名前はまだ無い。

私は猫である。名前はまだ無い。

→「吾輩」と「私」は同じような意味だ！

4. 類似度算出

文章レベルでベクトル化を行い，ベクトル同士のcos類似度を算出することで文章間の意味比較を目指す。

文章Aと文章Bのベクトルを$\boldsymbol{A}, \boldsymbol{B}$とするとcos類似度は

$$\cos\theta = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}$$

4. 類似度算出

「あんたバカ」を「あなたは馬鹿ですか」に置換した
アスカ訳.txt と アスカ.txtをあわせたものを学習データとした。

手法1:

Doc2Vecによる文章のベクトル化
Doc2Vecのメソッドを使用したcos類似度計算

手法2:　(精度が低かったので今回は未掲載)

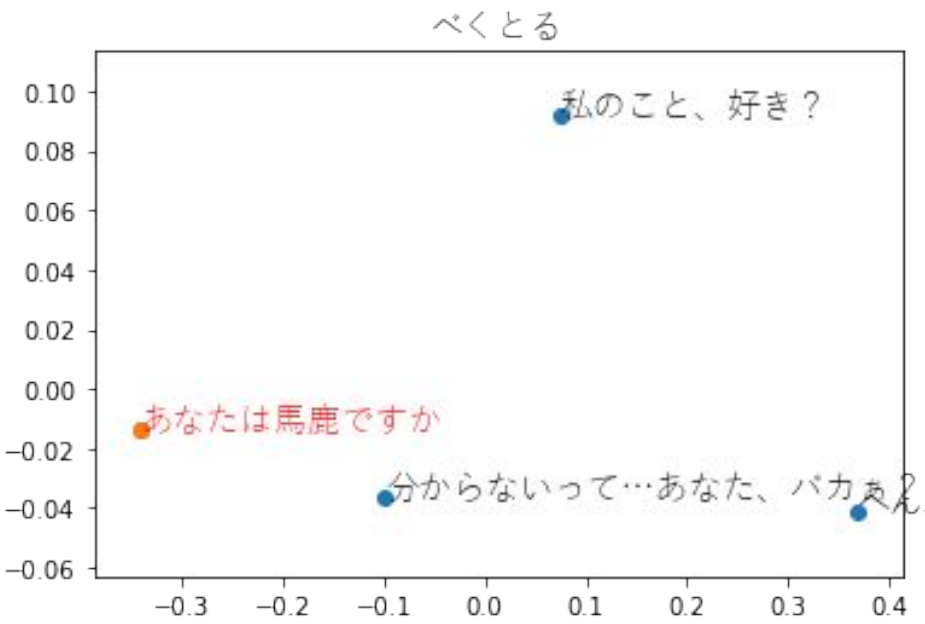
Word2Vecによる単語のベクトル化
単語ベクトルの平均をその文章の特徴ベクトルとして代用
自作のcos類似度モジュールで計算

4. 類似度計算

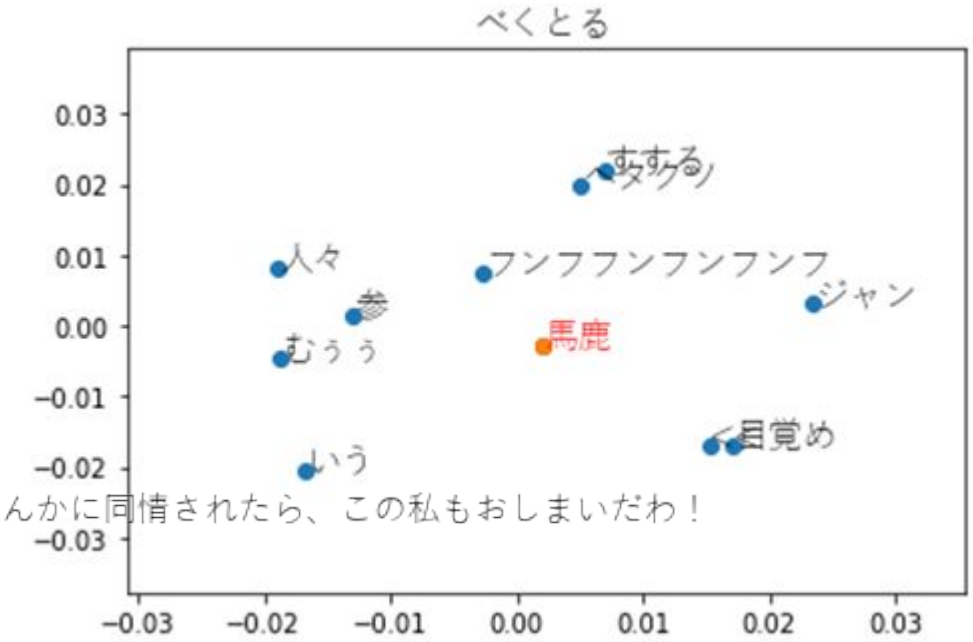
ちょっと練習...

Doc2vecで学習データからモデルを生成(分散表現)した後，

学習データ内の「バカ」と「あなたは馬鹿ですか」に類似する単語，文章を探してみた



「あなたは馬鹿ですか」に類似する文章



「馬鹿」に類似する単語

4. 類似度計算

単語レベルのLSTMで自動生成した文章(1万文)の中で
「あなたは馬鹿ですか」に類似する文章

Top 10

('お、これ！模試にられ。', 0.98114791898435283)
('父親作戦じゃないよう', 0.9769727574128275 2)
('いやてる～よ。渡すだっ！', 0.9729118888090361 2)
('あんたバカ、与えなのね', 0.9726307125024460 4)
('…私は、倍の時がいい', 0.9716968638338057 1)
('あの時にえぇっ', 0.9708652377544518 8)
('失礼てる…', 0.9691014852190796 8)
('あんたバカぁ？緊急の？', 0.9678645150179514 5)
('ママ！も私零中ねぇ。', 0.9659173202362962 8)
('あの女になんかの？', 0.9654918012304701 8)

“あんたバカ”系列の文

('あんたバカ？か…', 0.9203856895353237 4
('あんたバカじゃない！？', 0.8859560706280249 5)
('あんたバカ…よ。', 0.885878138144821 3)
('あんたバカの？', 0.8760652621233454 6)
('あんたバカ！', 0.859259088001117 06)

6. まとめ

○LSTMによる自動生成

・アスカの台詞は会話文の片側のみ抽出しているため文同士の関連度が低い。
・学習データにない単語，文字は生成できない。
・学習データに複数スタイルの文章を用いることはできない。

○Doc2Vecによる類似度

・アスカの台詞は文同士の関連度が低い
・用意したデータセットが間違い？

○今後の展望

・複数スタイルの学習データで学習して，実行時にスタイルの選択ができるようなseq2seqのモデルを作りたい。GANやVAEも加えられたらなお良し

・ライブラリ頼りの実装なのである程度自分の力で書いてみたい

・LSTMやDoc2Vecの理解を深めたい。

7. Appendix

プロ研を通しての感想/振り返り

▽良かったところ
プログラミングで何かを自由に作る体験が初めてだったので、完成したときに達成感があり、今後の自信に繋がると思う。また、サンプルのコードを見ながらでもある程度動くものが作れることがわかり、機械学習へのハードルが自分の中で下がった気がする。

▽難しかったところ
プロジェクトの初めにテーマを決める際に 何をすればいいのか/何がしたいのか/何ができるか が分からなかったので難しかった。取り組むことへの見通しの悪さと、動きながら軌道修正していくのが大切だと感じた。

▽今後に向けて
自然言語処理について理解を深めることができてよかった。まだよくわかってない部分が多いが、この分野への興味がさらに強まったので今後も学んでいきたい。