1T-02

機械学習による文章生成

深井 宏剛　　松澤 智史

東京理科大学 理工学部 情報科学科

1. 研究背景

雑談とは人間の高度な知的活動である. 近年, 人工知能分野の発展は著しく, 以前は難しいとされていた自然言語の意味的理解も可能であり, 実用化も進んでいる. しかし, 雑談に関する人工知能分野の研究は発展途上にある. 雑談を行う人工知能に関する研究は, 日本において孤独は社会問題であり, その問題を解決する手段になる. また, 対話に対する様々な角度からアプローチすることは, 人間の知的活動の過程を明確化でき, 意義深い.

2. 研究の目的

本研究では, 提案する文章生成システムとの対話自体を継続し楽しむことを重視するため, 雑談に着目する. このシステムは, 対話文を生成する過程において, 多量な文章の学習により単語の意味情報をベクトル化したものを用いることで, 豊富な語彙に対する対話を可能とする. また, このシステムとの対話が, 人間とする雑談と比べて, 違和感がないか, 対話を継続したいかについて, 実験と評価を行う.

3. 基礎知識

3.1 形態素解析

形態素解析 [1] とは, 文法的な情報の注記がない自然言語のテキストデータを, 対象言語の文法や, 辞書と呼ばれる単語の品詞や活用などの情報に基づき, 意味を持つ最小単位に分割することである.

3.2 機械学習

3.2.1 ニューロンモデル

ニューロンモデルとは, McCulloch と Pitts によって考案された, 生体ニューロンの動作を単純化してモデル化したものである. ニューロンモデルの例を図 1 に示す.

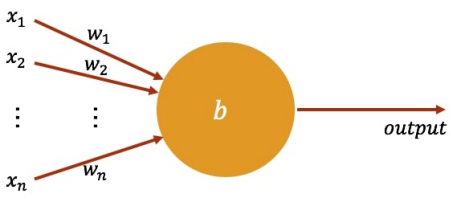

図 1: ニューロンモデルの例

このニューロンモデルは, 式 1,2 に示す数式に基づいて動作を行う.

$$net = \sum_{i=1}^{n} x_i w_i + b \tag{1}$$

$$output = f(net) \tag{2}$$

$x_i (i = 1, 2, ..., n)$ はニューロンモデルへの入力を表し, $output$ はニューロンモデルの出力を表す. $w_i (i = 1, 2, ..., n)$ はシナプスの結合効率を表す量で, 重みと呼ばれる. ニューロンモデルでは各入力に対して個別に重みを掛け合わせ, それらの総和とバイアス b を f に代入した出力を $output$ とする. f は活性化関数と呼ばれ, 用途に応じて様々な関数が用いられる.

3.2.2 ニューラルネットワーク

ニューラルネットワークとは, 人間の脳の神経回路構造を模した計算モデルである. 複数のニューロンモデルを結合することで構成され, ニューロンモデル間の重みを変化させることにより計算機に学習能力を与える. 例を図 2 に示す.

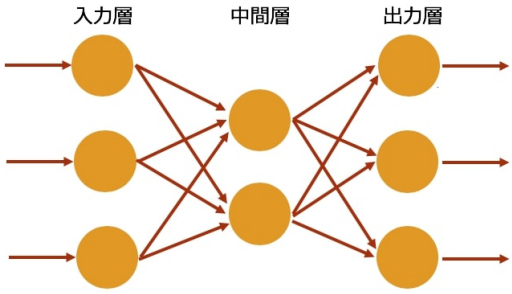

図 2: ニューラルネットワークの例

各層にはニューロンモデルが並んでおり, 受け取った入力を次の層のニューロンモデルに渡していき, 出力層から最終的な出力を得る. このような構造のものは特に順伝播型ニューラルネットワークと呼ばれ, 入力に対して出力が一意に定まるという特徴がある.

3.2.3 長・短期記憶ネットワーク

長・短期記憶 (Long Short-Term Memory,LSTM, 以下 LSTM という) とは, 再起型ニューラルネットワークの中間層内の各ユニットをメモリユニットとよばれる

The text generation by machine learning
Hirotaka Fukai,Tomofumi Matsuzawa
Faculty of Science and Technology, Dept.of Information Sciences, Tokyo University of Science

要素で置き変え, 入力を記憶するかリセットするかを決定する忘却ゲートを持った構造をもつニューラルネットワークである. 通常の再起型ニューラルネットワークでは理論上時系列データの入力全てを出力に反映できるが, その分中間層が深くなっていくことにより, 勾配の急激な増加もしくは消失となりやすい. その対策としての忘却ゲートの導入により, 重要な要素を長期的に残すことが可能となっている.

3.3 Word2Vec

Word2Vec[2] とは, Tomas Mikolov らによって提案された, 単語をベクトル表現するための手法である. 学習には, 2 層からなる順伝播型ニューラルネットワークを用いる. 学習により獲得されたベクトルは単語の意味情報を保持しており, 同一ベクトル空間内において似た意味の単語同士のベクトルは近い位置に配置される. 学習方法には, 出力層における周辺単語予測のエラー率の合計を最小化して推測する Skip-grams モデルを用いる.

4. 提案手法

4.1 Word2Vec による学習

日本語版 wikipedia の全記事を MeCab による形態素解析器を用いて形態素解析と分かち書きを行う. その後,Word2Vec により学習を行った. 提案手法の Word2Vec のパラメータを表 1 に示す. このパラメータで学習することにより, 228956 語の語彙と, その分散表現を獲得した.

表 1: 提案手法の Word2Vec のパラメータ

モデル	Skip-grams モデル
次元数	200
文脈窓	50
単語最低出現回数	50

4.2 LSTM による学習

対話文のデータセットを,Word2Vec の学習済みモデルに入力することで分散表現として獲得する. その後, その分散表現を LSTM により学習を行った. 提案手法の LSTM のパラメータを表 2 に示す. このパラメータで学習することにより, 対話文の分散表現を獲得した.

表 2: 提案手法の LSTM モデルのパラメータ

ニューラルネットワーク	LSTM
訓練データ	八割
テストデータ	二割
出力関数	softmax
最適化手法	RMSprop
損失関数	categorical_crossentropy
学習率	0.01
バッチ数	128
エポック数	60

4.3 対話文生成

Word2Vec の学習済みモデルによって, 入力文の分散表現を獲得する. その後,LSTM の学習済みモデルにその分散表現を入力することで, 返答文における形態素の分散表現を獲得する. それらの分散表現を Word2Vec の学習済みモデルに入力し, 形態素に変換して, 文章を生成する.

5. 実験

5.1 実験方法

システム評価者は, 挨拶文, 会話文, 単語などの様々な文章を入力する. システムが返答文を出力した際, システム評価者は同様の文章を入力する. 以上を繰り返し, システムとの対話を行う.

5.2 評価方法

システム評価者は, アンケートによる評価を行う. 項目として、「対話の妥当性」,「対話の継続性」を五段階評価で回答する.

5.3 結果

提案システムと対話した例を表 3 に示す.

表 3: 対話例

入力文	返答文
もういいのよ、 何でもしゃべって	U さんから 話さなくっちゃ
うん	あっ、じゃ、これ

6. まとめ

本研究では,Word2Vec と LSTM を用いることで,228956 語の語彙を持つ文章生成システムを構成した.

手法として, まず, ユーザから入力された文章を MeCab による形態素解析, 分かち書きを行う. そして, 得られた形態素を Word2Vec に入力することで, 文章の分散表現を獲得する. 対話文の学習をさせた LSTM モデルに, その分散表現を入力することで, 返答文を生成させる.

その結果, 提案した文章生成システムで雑談が可能であることを示せた.

参考文献

[1] Taku Kudo,Kaoru Yammamoto,Yuji Matsumoto
"Appllying Conditional Random Fields to Japanese Morphological Analysis"
Proc. of EMNLP, pp.230-237, 2004

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
"Distributed Representations of Words and Phrases and their Compositionality"
Advances in Neural Information Processing Systems 26, 2013