Vol.2016-MPS-107 No.1
2016/3/8

Deep Learning を用いた SNS からの知識抽出の一手法

鈴木由喜† 石川由羽† 髙田雅美† 城和貴†

本稿では，インターネット上の情報から集合知を抽出するために，Deep Learning を用いてクラスタリングし，知識の集約を行う一手法を提案する．インターネット技術の発達により SNS が盛んになり様々な知識が溢れている．その中には意図も根拠もない虚偽情報も多く含まれているため，膨大な知識から多様性，独立性，分散性を担保した集合知を抽出するのは困難である．そこで，Deep Learning を用いて SNS からのテキストデータを教師なし学習させ，クラスタリングをすることによって，集合知を抽出するために必要な多様性，分散性，独立性を担保した知識を集約する．

キーワード： 集合知，Deep Learning，SNS

A knowledge extraction method from SNS using Deep Learning

YUKI SUZUKI† YU ISHIKAWA†
MASAMI TAKATA† KAZUKI JOE†

In this paper, we propose a method for aggregating the knowledge by clustering using Deep Learning to extract the collective intelligence of information on the Internet. Through Internet technology is advanced, SNS is full of a variety of knowledge. Since it included many false information, it is difficult to extract the collective intelligence that fulfilled diversity, independence and dispersibility. To extraction collective intelligence with satisfied diversity, dispersibility and independence from SNS text data, unsupervised learning with Deep Learning should be adopted.

1. はじめに

インターネット技術の発達，ブログや SNS(Social Networking Service)サイトの普及により，様々な知識が世界中に溢れている．この膨大な知識の活用法として集合知の考えが取り入れられている．集合知の本来の意味は，群れのなかに宿る知のことであり，ネット上では，インターネットを利用して他人同士が知恵を出し合って知識を構築し，また利用することを意味する[1]． Web2.0 ではウェブが従来のウェブ(Web1.0)と異なる 7 つの原則を持ち，そのうちの 1 つに 2004 年にスロウィッキー(James Michael Surowiecki)が提唱した集合知を活用する．この Web2.0 は，2005 年にオライリー(Tim O'Reilly)が提唱した概念である．Web2.0 の潮流の中で特徴的な点は，ブログや SNS のような，参加者自身によるコンテンツの作成・公開である．ブログや SNS は，コンピュータの専門家だけでなく他分野の専門家，そして一般のユーザに対して情報公開の門戸を開いた．このような Web における集合知が専門家の持つ知識とは異なった価値を持つものであるとの主張がされている．以上の経緯から現代社会に集合知という言葉が受け入れられるようになる[2][3]．しかし，インターネット上には集合知のみではなく，炎上，噂，デマのように意図も根拠もない虚偽情報(集合愚)も多く生まれる．知識をただ集約し，利用しても集合愚が含まれる可能性が高く信頼性に欠ける．そこで集合知を抽出する必要がある．

集合知を抽出するには 3 つの条件(多様性，独立性，分散性)を満たす知識を集約する必要があり，これらが欠けると集約された知識は集合愚になりうる[2]．3 つの条件を満たす知識を集約した集合知は，専門家の知識より優れた知力を発揮する．集合知に関する既存研究としては，写真に付加されたユーザタグを集合知として用いてユーザの感覚に合致した地理情報を抽出する研究や，ソーシャルブックマークを集合知として用いて Web ページ推薦システムを構築する研究[5]がされている．しかし，これらの研究で使われている集合知は 3 つの条件を考慮していない．また，手動で集合知を抽出するにはビックデータを 1 つ 1 つ確認して 3 つの条件を満たすものを探さなければならないため，コストがかかる．したがって，自動的に集合知を抽出するシステムが必要となる．そのため，様々な人が利用する SNS をビックデータとして使用する．

SNS から大量のデータをラベル付けするには手間が必要であり，知識の正解不正解の決定には独断と偏見が伴うため，教師あり学習は好ましくない．さらに，大量のデータを扱うには高次元のベクトルを対象とする可能性が高い．高次元データを扱い，手動では難しい集合知の抽出を自動的に行うためには次元圧縮を伴い，教師なし学習で行う必要がある．Deep Learning は次元圧縮も行うことができ，データの抽象的な表現も得ることができるため，大量のデー

† 奈良女子大学
Nara Women's University

タをテキストマイニングし，知識の抽出が可能であると考えられる．以上の理由から，教師なし学習により，知識を学習する Deep Learning を適用する[6].

自然言語処理における Deep Learning を用いた既存研究はいくつかある[7][8][9]. しかし，ニューラル言語モデル[7]や畳み込みニューラルネット[8]には教師あり学習が必要であり，再帰 Autoencoder を適用する[9]には文の構造に依存してしまうため，ビックデータを扱うには適さない.

したがって，本稿では SNS から 3 つの条件を満たす知識を集約し，集合知を抽出するために，文構造に依存しない教師なし学習である Autoencoder の LSTM(Long short-term memory)版を用いて次元圧縮し，クラスタリングをする.

以下，2 章では，集合知について述べる．3 章では，提案する Deep Learning を用いた知識の抽出手法について説明する. 4 章では, 3 章で提案した手法を用いて実験を行う.

2. 集合知

2004 年にスロウィッキー(James Michael Surowiecki)が”The Wisdom of Crowds”(群衆の英知)で，3 つの条件を満たす知識を集約することで，集合知は専門家の知識より優れた知力を発揮すると主張する[10]．集合知が専門家の知識より優れた知力を発揮する例を 2 つ挙げる．1 つ目はテレビのクイズ番組『百万長者になりたい人は？』(日本版が『クイズ$ミリオネア』)である．回答者が答えがわからない場合，知人の中で最も知力の高い人に答えを訊くか，スタジオの視聴者にアンケートをとる．この場合，知力の高い人が適切であると考えられる．しかしながら，実際には知力の高い人の正答率が約 65 パーセントの正答率に対し，スタジオ観覧者のアンケートの正答率は 91 パーセントである．2 つ目の例がスペースシャトル・チャレンジャー号の爆発である．1986 年 1 月 28 日午前 11 時 38 分，スペースシャトル･チャレンジャー号が発射され 74 秒後爆発する. 爆発から 8 分後，最初の報道があり，間もなく株式市場では主要企業 4 社の株の投売りが始まる．その後，爆発から 21 分後，サイオコール社の株が突出して下落する．これは株式市場がチャレンジャー爆発の原因はサイオコール社にあることを意味する．爆発から 6 か月後，チャレンジャー号大統領調査委員会はサイオコールの部品が原因であると明らかにする．上記 2 つの例では集合知に必要な 3 つの条件である多様性，独立性，分散性を満たした知識を集約できたため，専門家の知識より優れた知力を発揮する．最初の例では様々な年代の視聴者(多様性)が各々アンケートに答えるため(独立性，分散性)，条件を満たすと考えられる．2 つ目の例でもある程度の規模の集団(多様性)が予測を行い，各々がそれぞれの場所(分散性)で回答を出す(独立性)ため条件を満たすと考えられる.

集合知に必要な 3 つの条件である多様性，独立性，分散性の定義は以下の通りである.

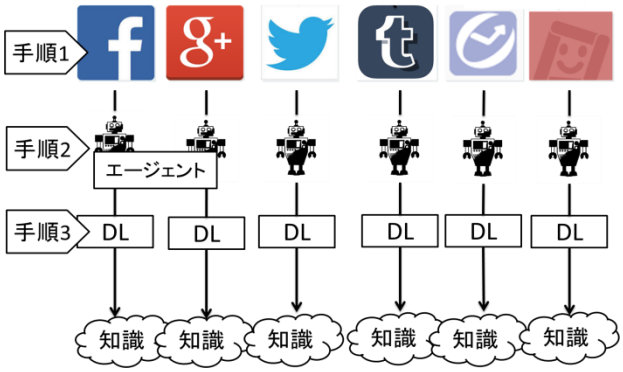

図 1　集合知の抽出手法

- 多様性：認知の多様性
- 独立性：他者の考えに左右されないこと
- 分散性：自立分散を意味し多様性や独立性をもたらすもの

多様性とは答えの候補の選択肢を増やし，新しい視点から検証できることである．個人の判断では正確さがなく一貫していないため，優れた意思決定には多様性が必要となる．独立性とは間違いが相互にかかわりを持たないようにできるほか，新しい情報を手に入れる可能性を持つことである．人は不安がある場合や状況が曖昧になると周りと同じ行動をする習性がある．その中で，嘘に流されないように意見の独立性が必要となる．分散性とはリーダーや上司に口出しされず，知の一部に対して自由に表現できることである. オープンソース OS の Linux や Wikipedia は高度な分散性を持つ集合知の成果である．3 つの条件を適切に集約しなければ，その知識は集合愚になることもある[2].

本稿の研究対象である SNS はインターネット技術によりどこからでも発信できるため分散性があり，様々な人が投稿するため多様性もあると考えられる．しかし，SNS にはフォローや RT(リツイート)，シェアなどがあり，他者の投稿に影響されやすいため独立性が担保できない．そのため，1 つの SNS を 1 つのコミュニティとする．そうすることでコミュニティ間は独立することになり，1 つのコミュニティから知識を抽出することで多様な知識が生まれ，かつ独立性が担保できると考えられる.

3. 集合知の抽出手法

3.1 手法の流れ

SNS からテキストデータを収集し，Deep Learning でクラスタリングを行うことで，多様性，独立性，分散性を満たす知識を抽出する．知識を集約し，集合知を抽出する手法の手順は，以下の通りである．図 1 は集合知の抽出手法を図に表したものであり，図中の DL は Deep Learning を略し

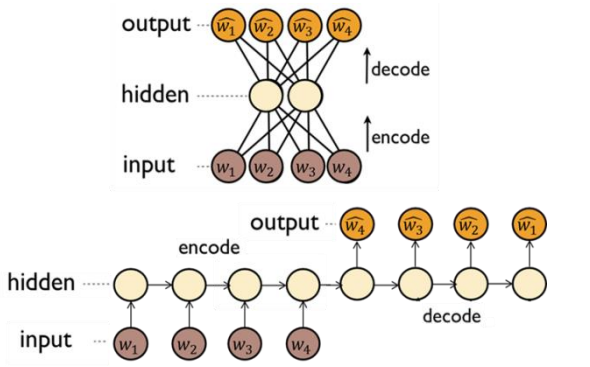

図 2　Autoencoder と LSTM 版 Autoencoder の違い

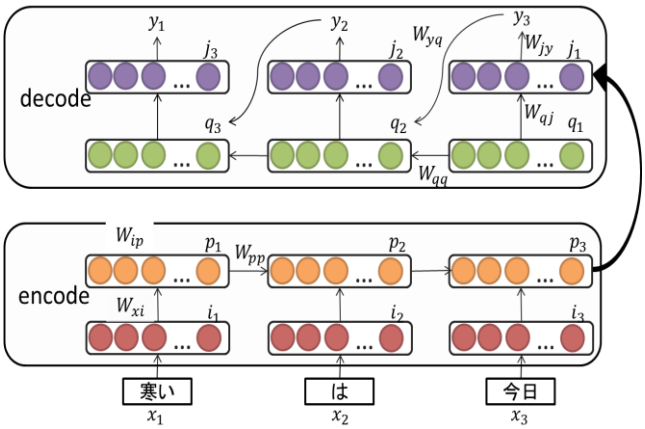

図 3　LSTM の概要図

たものである．

手順 1　知りたい知識に関するキーワードの設定
手順 2　マルチエージェントを用いたデータ収集・入力データの作成
手順 3　Deep Learning を用いたクラスタリング

手順 1 では，集合知の対象となるキーワードを設定する．例えば，「風邪」について知りたいとき，「風邪」をキーワードに設定する．手順 2 では，手順 1 で設定したキーワードに関するテキストデータを各 SNS から収集し入力データを作成する．詳細は，3.2 節で述べる．手順 3 では，Deep Learning を用いてクラスタリングを行う．Deep Learning を用いることで知識を抽出し，知識におけるクラスタリングができると考えられる．詳細は 3.3 節で述べる．手動では困難である集合知の抽出を Deep Learning とマルチエージェントを用いて自動で行うことができると考えられる[11].

集合知に必要な 3 つの条件は以下のように満たすことができると考えられる．

- 多様性：SNS からさまざまな人が投稿する，かつ類似した知識はひとまとめにされるため満たされる
- 独立性：1 つの SNS を 1 つのコミュニティとみなし，その独立したコミュニティから知識を集約するため満たされる
- 分散性：さまざまな人が各自分散して投稿するため満たされる

text(1)　t5 t3 t4 t1
text(2)　t4 t2

	t1	t2	t3	t4	t5
index	1	2	3	4	5

text(1) t5	5	0	0	0
text(1) t3	0	3	0	0
text(1) t4	0	0	4	0
text(1) t1	0	0	0	1
text(2) t4	4	0	0	0
text(2) t2	0	2	0	0

図 4　単語ベクトルの生成

3.2 マルチエージェントを用いたデータ収集・入力データの作成

本節では，SNS からテキストを収集し，入力データを作成するまでの手順を説明する．

手順 2 - 1　マルチエージェントを用いた SNS からのテキストの収集
手順 2 - 2　テキストのわかち書き
手順 2 - 3　入力層に入力

手順 2-1 では，SNS からマルチエージェントを用いてテキストを収集する．SNS からテキストを収集するには周期的に取得するクローラを使用する場合がある．クローラとはリンクをたどって情報を自動収集するウェブページ探索プログラムのことである[2]．しかしクローラを用いると，ある SNS で話題になっている事柄が別の SNS ではどのように話題にされているか判断するのが難しい．さらに，小さな SNS(コミュニティ)自体を取得するのも困難である．マルチエージェントを用いることで以上の点が可能になると考えられる[12]．ただし，本稿では，Deep Learning で知識を抽出することに重点を置くため，マルチエージェントは用いず手動でデータ収集を行う．SNS を絞り，その SNS 上で知識を抽出できるかを検証したのちに，マルチエージェントを用いたいと考えている．手順 2-2 では，既存の形態素解析エンジン[13]を用いて，テキストをわかち書きにする．「今日は寒い」というテキストデータがあるとすると，「今日 は 寒い」となる．手順 2-3 では，テキストデ

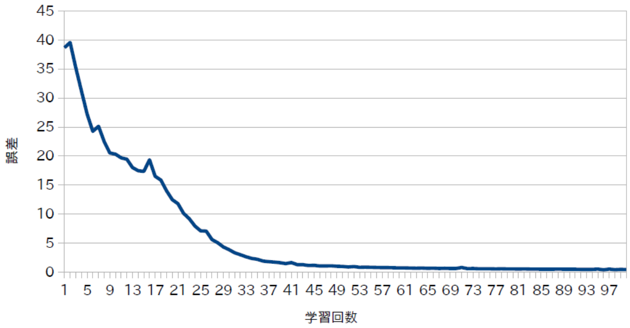

図 5 DL の encode と decode の誤差の推移

ータを逆順に入力する．逆順に読み込むことで文頭の情報をできるだけ残す．これにより，短期依存関係の問題を解決し，最適化がはるかに容易になる[14].

3.3 Deep Learning を用いたクラスタリングの手順

本稿は Nitish らの研究[15]を参考にして Deep Learning を用いたクラスタリングを行う. この研究は, Autoencoder[16]を LSTM[14]に適用させたもので，ビデオを時系列の画像データとして学習させている. LSTM とは, encode と decode の 2 個の RNN(Recurrent Neural Network)を用意し，中間ノードで繋ぎ合わせたものである．図 2 に元の Autoencoder と LSTM 版 Autoencoder の違いを示す．図 2 の上図が元の Autoencoder であり，入力データx_1~x_4の特徴を encode で学習し decode でデータの再現を行うものである. 入力データx_1~x_4と出力データy_1~y_4の差を最適化しながら学習する．下図は LSTM に応用させた Autoencoder であり，入力データx_1~x_4を時系列に encode で学習し，decode で時系列に再現する．どちらも教師なし学習であるが，LSTM を用いることで時系列に対応した学習ができる．図 3 に LSTM 版 Autoencoder の概要図を示す．$\boldsymbol{x}$は入力データ，$\boldsymbol{y}$は出力データ，$\boldsymbol{i}$と$\boldsymbol{j}$はそれぞれ入力データと出力データの埋め込み層のユニット，$\boldsymbol{p}$と$\boldsymbol{q}$はそれぞれ encode と decode の隠れ層のユニットである．W_{xi}，W_{ip}，W_{pp}，W_{qq}，W_{qj}，W_{yq}はそれぞれの箇所への重みである．encode で入力データを学習を終えると最後の隠れ層をコピーし decode で再現を行う．Deep Learning を用いたクラスタリングの手順を以下に示す.

手順 3 - 1 単語ベクトルに変換し埋め込み層に入力
手順 3 - 2 単語ベクトルを 1 つ 1 つ encode
手順 3 - 3 単語ベクトルの数だけ decode
手順 3 - 4 Affinity Propagation を用いてクラスタリング

手順 3-1 では，学習する文書データを単語ベクトルに変換する．単語ベクトルの変換の例を図 4 に示す．テキストデータが 2 つ(text(1),text(2))，出現単語が 5 語(t_1~t_5)ある場合，2 つのテキストデータの長い方(text(1))に合わせた配列

表 1 クラスタリング結果

一致した単語の語数	クラスタ数
1語	33
2語	2
3語	1
その他	77
合計	113

表 2 クラスタリング結果の内容の一例

クラスタ番号	クラスタの内容
7	• やべぇー風邪ひいた
10	• やべー喉がいたいこれは風邪引いた
32	• 完全に風邪ひいただせえ • 完全に、風邪だ、これは、やばい、です、ウワッ、ねよ、、 • 完全に風邪ひいたな?喉痛いし鼻水止まらん。でもだるくはない！あとは屋根組んで終わりだ • 完全に風邪 • 完全に風邪でわろた。帰りに薬買って帰ろう。。。 • 完全風邪だ。
78	• 昨日の変な天気のせいで風邪ひいたかも • 昨日、一昨日と久しぶりの連休。天気も良かったが何もせず、ただ引きこもってた2日間、なのに今朝起きたら身体は怠いゎ、のどは痛いゎ、完璧な風邪症状…….。(;>_<;)ﾏｲｯﾀ! • 昨日口を開けて寝ていたのか朝からずっと喉が痛いんだが風邪…？風邪かも…？風邪かもしれない！(ﾟﾛﾟ麻生くん風に読んでみよう！) • 昨日の私「風邪？何年引いてないと思ってんの wwwwwww かかるわけないじゃん wwwwww」今日の私「見事に風邪引いた…」素晴らしいフラグ回収でした。 • 昨日は(･ム･)とkiyoちゃんがお出かけしてソッコー(^-^)がツイしてきたんだょね…wそれってﾔｷﾓﾁあぁ、風邪っぴきのﾎﾞｸの心配してくれないの？(^-^;)とかいうﾔｷﾓﾁ • 昨日、今日と長男学校休みおって病院連れて来てます。男の子って弱い…私を見てみなさい！風邪も寄って来ないよ(´･ω･) • 昨日の余韻に浸る余裕もなく風邪の諸症状と闘い中寒気は治まったから熱は上がりきったかな？ • 昨日病院でついでに貰った風邪薬が副作用欄の項目多すぎて、飲むのをビビる • 昨日インフルエンザの注射したのに、今日風邪ひいたわw • 昨日インフルエンザの予防接種に行ってきたんですが、私の場合は微熱・だるさ・眠気と風邪に似た症状が出ました。1日たった今はだるさと眠気がちょっと残ってます。
82	• この私が風邪をひいた • この私が風邪をひくなんてありえない • この前、風邪のひき始めでワインがんがん飲んで良くなったかも、いけるか！って思ってたのは完全にアルコールで麻痺してたせいだったんだな。いけてないです。 • この季節になるとよくあることだが、「風邪ひいちゃったけど頑張って来たー！(・∀・)v…げほっげほっ…ぶぇっっくしょい！！」というのはマジでテロなのでやめて欲しい…学生でも先生でも… • この会社フロア毎で気温差5度以上あるんだけど…絶対風邪ひくべ。
89	• 今夜はお家で飲もう(.???.)やらないで！やめて！はやく処理して！って言ったのに約束守ってくれなかった営業全員風邪拗らせろ(・∀・)！ • すいません。。。ガチ風邪です。正確には風邪じゃないんですけどウイルス性の病気です。。。。今、頭痛、吐き気、お腹です。。。。。 • 子供は熱出すし(元気だけど)怪我もしてるし(元気だけど)自分も風邪気味だし久々の3歳児クラスも休まなきゃだし出かける予定あったのに残念。昨日3時間残業したから疲れてたし良かったのかな?子供がいる前提での生活は自由ではないよね。当たり前だけど、もどかしい気持ちはあるよ。 • 乙哉「コンビニ寄っていい？」しえな「なんで？」乙哉「しえなちゃん風邪気味でしょ？マスク買ってくる」しえな「あ、ありがと…///」乙哉「おまたせ」しえな「えっなんでマスクつけてんの」乙哉「え、だってしえなちゃん風邪気味でしょ？」 • イルカも風邪をひく • うわーこれ完全に風邪だーー！！これ普通なら家で寝てるやつだー！！なんで働いてるんだよ！！！(人手不足) • 髪をすすいでるときシャワーを水にしたら反撃で水かけられた、風邪引いたらおねえちゃんに看病してもらうんだ(???) • 完璧風邪ひいた • また風邪ひいたようだ。かなり吐いた • また風邪ひいたかなまじめに(笑)気温差はげしいちゃ！(笑) • 来年こそは1回も風邪を引かずに健康に過ごします！(空) • また風邪ひいちゃった。。。 • ぶっちゃけなんで今風邪引いてんのかほんとクソだと思うんですよ。マスク嫌いなのにーーーー！！！ • やっと風邪が落ち着いた感編集作業に移るか
108	• おはようご財増す！今朝から風邪の初期症状のだるさを感じ吐き気を気にせずパンを食べて市販の薬をのんで散髪に来ています。帰ったら長男の七五三をやって明日から2日間の旅公演の為夕方から日立に向かいます。 • おはようございます^_^わたしは、体調くずしました。体調管理大事なので気をつけてくださいね^_^風邪や病気にならないように願います。 • おはようございます、昼過ぎから爆睡してました風邪治りました • おはようございますm(_ _)m風邪をひきました。今日は家で寝てます。(；＿；) • おはようございます^^喉痛いから風邪かなぁ。

を作り, 出現する場所にその単語のインデックスを入れる. 他は 0 パディングを行う．単語ベクトルに変換した後，次元を圧縮し埋め込み層$\boldsymbol{i}$に単語情報を表す. 式は以下のようになる.

$$i_{\mathrm{n}} = \tanh(W_{xi} \cdot x_n) \tag{1}$$

手順 3-2 では, 単語を 1 語ずつ LSTM に入力し学習させる. 隠れ層$\boldsymbol{p}$を表し，式は以下のようになる.

$$p_{\mathrm{n}} = \mathrm{LSTM}(W_{ip} \cdot i_n + W_{pp} \cdot p_{n-1}) \tag{2}$$

手順 3-3 では，すべての入力単語の学習が終わり次第，入力単語数と同数 decode させる．隠れ層$\boldsymbol{q}$，埋め込み層$\boldsymbol{j}$は，それぞれ以下の式になる.

$$q_{\mathrm{n}} = \mathrm{LSTM}(W_{yq} \cdot y_{n-1} + W_{qq} \cdot q_{n-1}) \quad (3)$$

$$j_{\mathrm{m}} = \tanh(W_{qj} \cdot q_n) \quad (4)$$

$$y_{\mathrm{m}} = W_{jy} \cdot j_n \quad (5)$$

入力のベクトルとデコードの結果のベクトルの最小平均二乗誤差を誤差関数とし(式(6))，誤差が小さくなるよう手順 3-1 から手順 3-3 までを繰り返し，最適化させながら学習させる．

$$\mathrm{argmin}\frac{1}{n}\sum_{i=1}^{n}\|y_i - x_i\|^2 \quad (6)$$

手順 3-4 では，学習が終わった encode の最後の隠れ層$\boldsymbol{p}$を教師なし学習である Affinity Propagation を用いてクラスタリングを行い，出力する．Affinity Propagation を用いる利点として，高次元のデータでもクラスタリングを行えること，クラスタ数をあらかじめ指定しなくてもよい点が挙げられる[17].

4. 実験と考察

4.1 環境

SNS から収集したデータを Deep Learning を用いてクラスタリングする. 本稿では Deep Learning を用いたクラスタリング結果の検討をするため，マルチエージェントを使わず，SNS から無作為にテキストデータを収集する．

クラスタリング後の結果を評価しやすくするために，キーワードを指定し，2015 年 12 月 8 日に収集したデータを使用する．本稿では，キーワードを「風邪」に設定し，SNS (Twitter[18]，はてなハイク[19])から収集された 594 件のデータを学習データとして使用する．

4.2 結果

Deep Learning の encode と decode の誤差の推移を図 5 に示す．学習回数が 100 回で誤差の推移が一定になった．誤差が低下することから学習できていることが確認できる．Affinity Propagation の結果，113 のクラスにクラスタリングされた．クラスタの中身を見ると，表 1 から先頭の単語に沿ってクラスタリングされていることがわかる．表 1 に先頭の単語が一致したクラスタ数を示す．先頭から一致した単語のクラスタ数はそれぞれ 1 語が 33，2 語が 2，3 語が 1 である．その他は，1 クラスタあたり 1 データしかないものや，内容が様々なものが多数入ったクラスタなどが含まれている．表 2 は 113 のうち特徴的な 7 つのクラスタの一例である．表 2 の左が番号，右が中身である．中身のテキストデータはわかち書きした後のデータである．クラスタ番号 7，10 は多数ある 1 文のみのクラスタのうちの 2 つである．先頭の単語が一致しなかったものがほとんど 1 文のみのクラスタに入っている. クラスタ番号 32 からは「完全」「に」「風邪」まで読み取れていることがわかる．「完全に風邪」と「完全風邪」を同じ意味だと捉えられている．クラスタ番号 78 のクラスタからは先頭の単語に着目したクラスタリングができている. クラスタ番号 82 のクラスタからは「この」に着目しつつも「この私が風邪」まで関連性があると捉えられていることがわかる．しかし「この」に着目したため他の単語の意味を捉えられず，意味の違うテキストデータも含まれている．また，クラスタ番号 108 のクラスタからわかるように，誤変換でも同じ意味(「おはようございます」と「おはようご財増す」)として捉えられている．

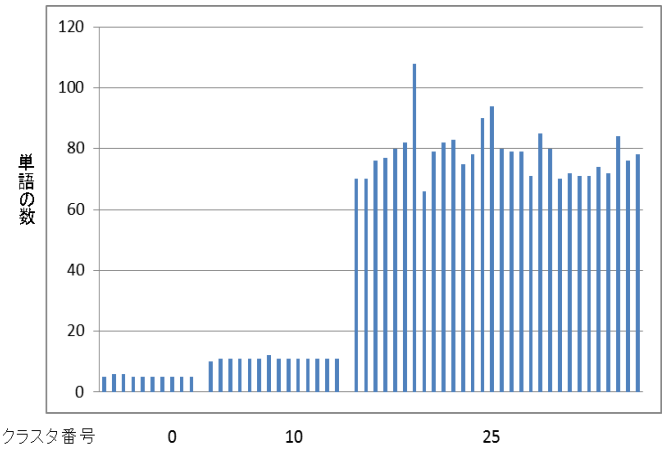

図 6　入力データをクラスタリングした結果の一例

4.3 考察

表 2 より，先頭の単語を学習しただけではなく「おはようございます」の意味も捉えられていること,「完全に風邪」と「完全風邪」が同じクラスタであることから，隠れ層ではテキストデータの単語間の意味を把握したベクトル表現ができたと考えられる. しかし, クラスタ番号 7 と 10 の「やべぇ」と「やべー」は同義語として認知されなかったため別々のクラスタに含まれたと考えられる．したがって，単語の意味まで把握できるとは言えない．図 6 は Deep Learning で学習させる前の状態で Affinity Propagation をかけた結果である. 短文のクラスタ, 中間の長さのクラスタ, 長文のクラスタの 3 つの横軸がテキストデータの数，縦軸が単語数である. 学習前データをクラスタリングすると 37 のクラスタに分かれたが，図 6 のようにテキストデータの長さに依存した結果になってしまい，意味を捉えることは不可能だった．Deep Learning を用いて学習させることで長さに関係なく語順を考慮し，文脈の意味も考慮したベクトル表現ができると考えられる．しかし本稿では先頭 1~2 語でクラスタリングされたものがほとんどであるため，実験の学習回数やユニット数などを検討することで学習する単語の範囲が増え，より詳細な意味表現ができるのではないかと考えられる．

関連したテキストデータが同じクラスタに入ることから，

Deep Learning を用いたクラスタリングは類似したデータをまとめることが可能である．Deep Learning を用いて大量のデータを学習させ，類似したデータを排除することで，集合知を抽出するにあたり，多様性の向上が行えると考えられる．分散性は SNS で各々の場所から投稿することからすでに満たされている．独立性はコミュニティ間が独立しているため，担保できていると考えられる．しかし SNS は人の影響を受けやすく独立性を担保することが難しいため今後の課題でもある．したがって，過去の集合知の知見を用いて SNS からテキストデータを収集し，実際に自動的にその集合知を抽出できるかを確認する必要がある．また多様性をさらに強化するために，SNS の数を増やし，マルチエージェントを用いることが好ましい．マルチエージェントを用いることで自動的に多数の SNS からテキストデータを収集すると考えられるからである．

5. まとめ

本稿では，Deep Learning を用いてクラスタリングし，知識を抽出するための一手法を提案した．インターネット技術の，ブログや SNS サイトの普及により，様々な知識が溢れている．この膨大な知識から 3 つの条件を満たした集合知を抽出する．3 つの条件とは多様性，独立性，分散性のことである．3 つの条件を満たさなければ，知識は集合愚になりうる. したがって, SNS から大量のデータを収集し，自動的に 3 つの条件を満たす集合知を抽出する必要がある. 集合知を抽出するために SNS からテキストを収集し, Deep Learning を用いてクラスタリングを行う．まず，SNS からテキストを収集し，単語を逆順に入力層に入力する．次に Deep Learning を用いて特徴抽出を行い, Affinity Propagation を用いてクラスタリングを行う．Deep Learning の一種である Autoencoder の LSTM 版を本稿では採用している．

SNS からキーワードを設定したテキストデータ 594 件を収集し，学習データとして用いて実験を行った．実験の結果より，語順に配慮した長さに依存しないクラスタリングができた．隠れ層ではテキストデータの単語間の意味を把握したベクトル表現ができた. この結果より, Deep Learning を用いたクラスタリングで類似したデータを排除し，集合知の多様性の強化を行うことができた．したがって，集合知に必要な 3 つの条件は満たされたと考えられる．多様性は，SNS から様々な人が投稿し，かつ類似した知識はひとまとめにされるため満たされる．独立性は，1 つの SNS を 1 つのコミュニティとみなし，その独立したコミュニティから知識を集約するため満たされる．分散性は，様々な人が各自分散して投稿するため満たされる．しかし未だ確証は得られないため，実際に過去の集合知を使って実験を行い，Deep Learning で知識を抽出できるかを検証する必要がある．

今後は，さらに意味表現の精度が上がるように学習係数や重みの初期値の設定を試行錯誤しなければならない．そして，本稿では手動で SNS からテキストを収集したため，マルチエージェントを用いて多数の SNS からテキストを収集する必要がある．

参考文献

[1] 西垣通:”集合知とは何か～ネット時代の「知」のゆくえ～”, 中央公論新社,2013.2.25
[2] 増永良文:”ソーシャルコンピューティング入門”,株式会社サイエンス社,2013.3.25
[3] 大向一輝” Web2.0 と集合知” 国立情報学研究所,情報処理 47(11), 1214-1221, 2006-11-15
[4] 末田航, 味八木崇, 暦本純, :”実世界集合知による利用者の認知地図の可視化とモバイルインタラクションへの適用” 情報処理学会論文誌 Vol.52 No.4 1465-1474 2011
[5] 丹羽 智史,土肥 拓生, 本位田 真一“Folksonomy マイニングに基づく Web ページ推薦システム” 情報処理学会論文誌 Vol47 No.5 2006
[6] 浅川伸一:”ﾃﾞｨｰﾌﾟﾗｰﾆﾝｸﾞ,ﾋﾞｯｸﾃﾞｰﾀ,機械学習あるいはその心理学”,新曜社,2015.2.10
[7] Collobert, R. and Weston, J.: A Unified Architecture for Natural Language Processing: Deep Neural Networks with MultitaskLearning, in ICML 2008, pp. 160 –167 (2008)
[8] Rie Johnson, Tong Zhang: Effective Use of Word Order for Text Categorization with Convolutional Neural Networks: arXiv preprint arXiv:1412.1058, 2014.
[9] Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., andManning, C. D.: Dynamic Pooling and Unfolding Recursive Autoencodersfor Paraphrase Detection, in NIPS’11 (2011)
[10] ジェームズ･スロウィッキー,小高尚子訳,”「みんなの意見」は正しい”,角川書店,2009.11.25
[11] 小林一郎:” 人工知能の基礎” ,株式会社サイエンス社,2008.11.10
[12] 伊藤孝行,金森亮,チャクラボルティ シャンタヌ,大塚孝信,原圭佑: 未来の社会システムを支えるマルチエージェントシステム研究(1)―経済パラダイム,交渉エージェント,交通マネジメント―, 人工知能学会誌 28 巻 3 号(2013 年 5 月)
[13] A WEB Page, Page,Kyoto University (online), available from <http://mecab.googlecode.com/svn/trunk/>, (accessed 2014-02-01).
[14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215, 2014.
[15] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. arXiv preprint arXiv:1502.04681, 2015.
[16] 岡谷貴之:”深層学習”,講談社,2015.4.17
[17] Brendan J. Frey and Delbert Dueck: “Clustering by Passing Messages Between Data Points”, Science, vol.315, pp.972-976(2007)
[18] Twitter API https://dev.twitter.com/ (accessed 2015-12-08).
[19] はてなハイク API http://developer.hatena.ne.jp/ja/documents/haiku (accessed 2015-12-08).