深層学習を利用した
レビュー文章分類モデルと特徴量の抽出

中央大学　理工学研究科　経営システム工学専攻
三宅　伸

電子商取引（Electronic Commerce，以下**EC**）市場の拡大

商品の購買やサービスの予約などが，手軽に行える

消費者の行動，生活環境に大きな変化をもたらした

EC市場の拡大に伴い，
レビューなどの**情報共有サービス**が普及傾向にある．

情報共有サービス

レビューの投稿

・消費者の体験に基づく評価
・顧客の購買決定を大きく左右する

➢ 顧客が実際に体感した内容を文章として表現するため，
投稿されたコンテンツを分析することは非常に重要である．

深層学習を用いたモデルの普及

コンピュータビジョンにおいて驚くべき成果をあげている

消費者が投稿したテキストデータを対象とした取組み
カテゴリ分類などの分類問題，分析やスパム検出およびその他

従来の自然言語処理タスクにおいて，深層学習を用いたモデルが有効であると考えられる.

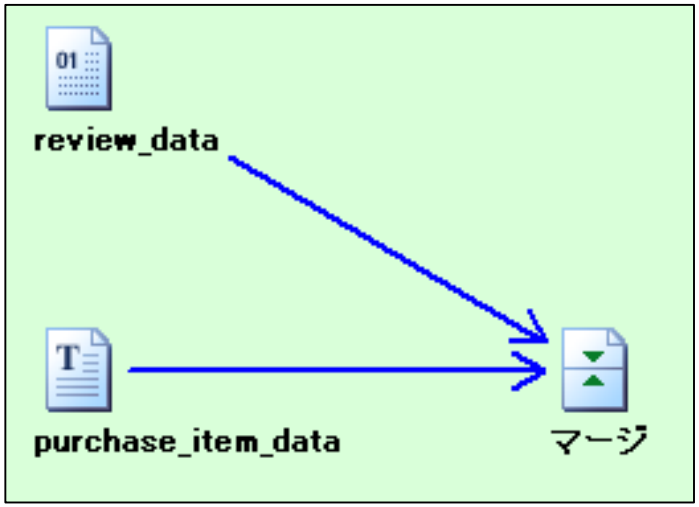
本研究では、
投稿者が実際に購入した商品のカテゴリを、投稿コンテンツを用いて判別する文章判別モデルを作成し、判別する。

また、作成した判別モデルの学習で、どのような単語が判別に有効であったのかを考察し、各カテゴリの特徴を考察する。

具体的には、レビュー文章を系列データとして、
リカレントニューラルネットワークの一種であるLong short-term memory（LSTM）による商品カテゴリの判別モデルを作成する。
学習に関する特徴的単語は、言葉ネットワークを利用して考察する。

分析データの概要

使用データ：ゴルフポータルサイトより提供
対象期間　：2013年11月01日　～　2016年11月30日
対象投稿者：期間内にゴルフのギア商品を購入している顧客
　　購入した商品に関するレビューを投稿している顧客



商品の購買データ
（EC内で扱っている、商品の購買データ）

ギア商品の購入データ

商品のIDと、レビュー対象の商品IDで
データをマージし、分析用のデータを作成

分析で利用するギア商品のカテゴリ（ラベル）

商品カテゴリ	レビュー件数
アイアン	650件
ウェッジ	**783件**
ドライバー	**900件**
パター	**712件**
フェアウェイウッド	571件
ユーティリティ	600件

商品カテゴリを文章のラベルとする.
なお、カテゴリ別で商品の投稿件数が多い
3カテゴリをモデルで利用する.

ウェッジ、ドライバー、パター

投稿されたレビューコンテンツ（学習データ）

レビューの内容を学習モデルで利用するためのデータの形成方法を
次のスライドにあらわす.

実際に投稿されたレビュー例

全体的な感想は、振り易く、バランスD3となっていますが重さを感じる事なく振れる。
確実に１番手は違う、セットの見直しが必要かもです。
デザインは　かっこいい、満足飛距離は、飛び過ぎ！
打感は　ハジキ感がすごいが固さは感じない。
方向性は　操作性は高い、普通に打つとヘッドの返りがよく捕まったドロー。
弾道の高さは　ストロングロフトだけど高弾道

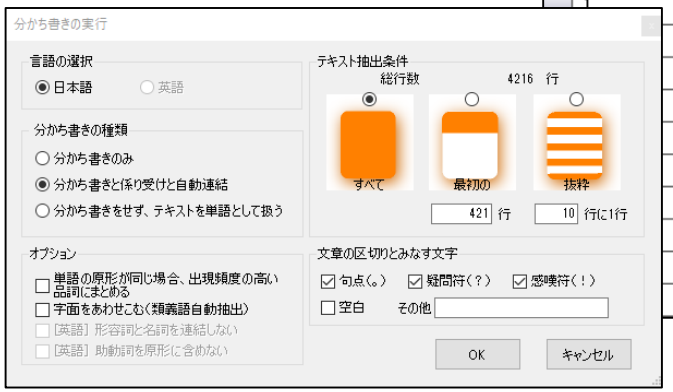
全体的な感想は少し軽く感じますデザインは良い飛距離は飛びます打感はまあまあ方向性は強いドロー(フック)弾道の高さはなかなか高い

→　３つのカテゴリに該当するレビューを選択し，**形態素解析**を行う

AMTとの重量の流れもよかたので購入しました。
練習場では今まで使用していたアイアンよりも、あきらかに距離が伸びています。
どれくらい距離がでるのかコース場で早く確認したいのです。

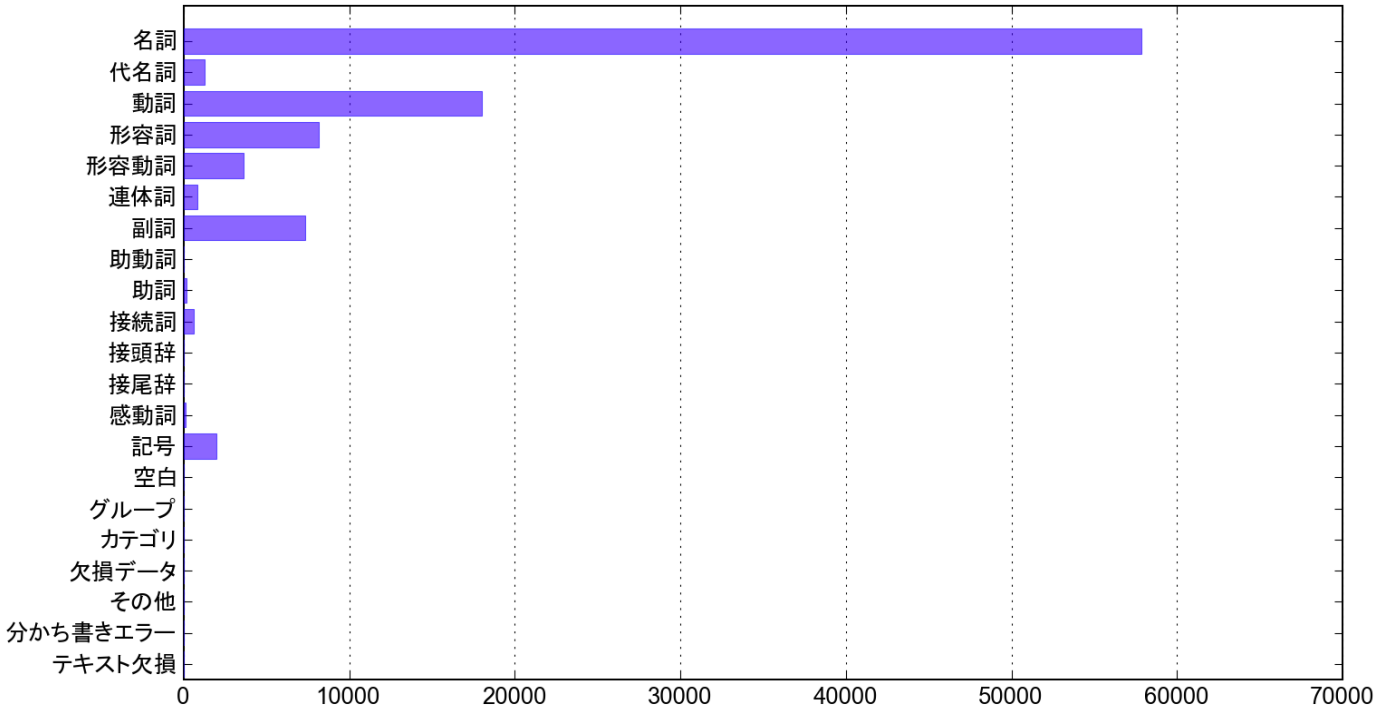
形態素解析

ファイルID	行ID	文章ID	単語ID	見出し語	原形	置換語	品詞	品詞詳
1	1	1	1	AMTとの	AMT	AMT	名詞	一般
1	1	1	2	重量の	重量	重量	名詞	一般
1	1	1	3	流れも	流れ	流れ	名詞	一般
1	1	1	4	よかたので	よかた	よかた	名詞	一般
1	1	1	5	購入しました。	購入	購入	名詞	サ変可能
1	2	1	1	練習場では	練習場	練習場	名詞	一般
1	2	1	2	今まで	今	今	名詞	副詞可能
1	2	1	3	使用していた	使用	使用	名詞	サ変可能
1	2	1	4	アイアンよりも、	アイアン	アイアン	名詞	一般
	2	1	5	あきらかに	あきらか	あきらか	形容動	一般
	2	1	6	距離が	距離	距離	名詞	一般
	2	1	7	伸びています。	伸びる	伸びる	動詞	一般
	3	1	1	どれくらい	どれ	どれ	代名詞	
	3	1	2	距離が	距離	距離	名詞	一般
	3	1	3	でるのか	でる	でる	動詞	一般
	3	1	4	コース場で	コース場	コース場	名詞	一般
	3	1	5	早く	早い	早い	形容詞	一般
	3	1	6	確認したいのです。	確認	確認	名詞	サ変可能



→ 形態素解析結果から，**名詞，形容詞，動詞**のみを選択

	品詞	出現回数
1	名詞	57864
2	代名詞	1274
3	動詞	18039
4	形容詞	8172
5	形容動詞	3598
6	連体詞	823
7	副詞	7360
8	助動詞	34
9	助詞	190
10	接続詞	598
11	接頭辞	2
12	接尾辞	5
13	感動詞	144
14	記号	1992
15	空白	6
16	グループ	0
17	カテゴリ	0
18	欠損データ	0
19	その他	0
20	分かち書きエ	0
21	テキスト欠損	0



上記の表は，レビュー内に含まれている品詞の分布を表す．
名詞，動詞，形容詞の順で頻出回数が多いことがわかる．

数値データへの変換例

レビュー１：
打感は柔らかく，ミスに強く，軽いフェード，軽いドローと操作性も良い
レビュー２：
飛距離に差が出ないため、非常に良いと思います．

[“打感”,”柔らかい”,”ミス”,”強い”,”軽い”,”フェード”,”軽い”,”ドロー”,”良い”]
[“飛距離”,”差”,”出ない”,”非常に”,”良い”,”思う”]

[1, 2, 3, 4, 5, 6, 5, 7, 8]
[9,10,11,12,8,13]

単語が出てきた順に，数値番号を付与する

[1, 2, 3, 4, 5, 6, 5, 7, 8]
[9,10,11,12,8,13]

N次元＝100次元に設定

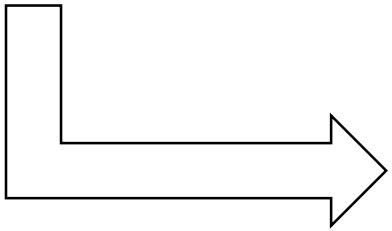
0	0	0	0	1	2	3	4	5	6	5	7	8
0	0	0	0	0	0	0	9	10	11	12	8	13

文章行列をパディングし，数値行列としてモデルに導入する.

→ 単語行列をシーケンスデータとして利用する
LSTMの特性上後の（時系列の後）後ろの単語の方が重要視されるため

[1, 2, 3, 4, 5, 6, 5, 7, 8]
[9,10,11,12,8,13]

N次元＝100次元に設定



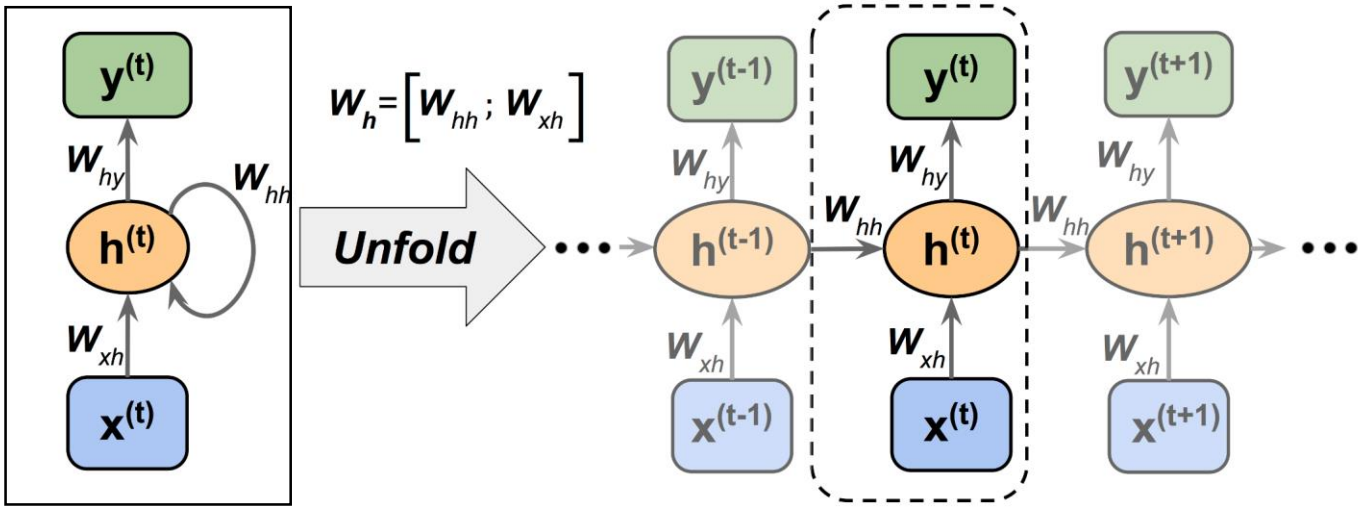
0	0	0	0	1	2	3	4	5	6	5	7	8
0	0	0	0	0	0	0	9	10	11	12	8	13

本研究では，センテンスの次元を100次元に設定
モデルの入力には，この数値化したレビューデータを代入する.
ラベルは３つのカテゴリから，３つに設定する.

なお、学習データに全体の7割を利用，テストデータに3割を利用する.

LSTMについて

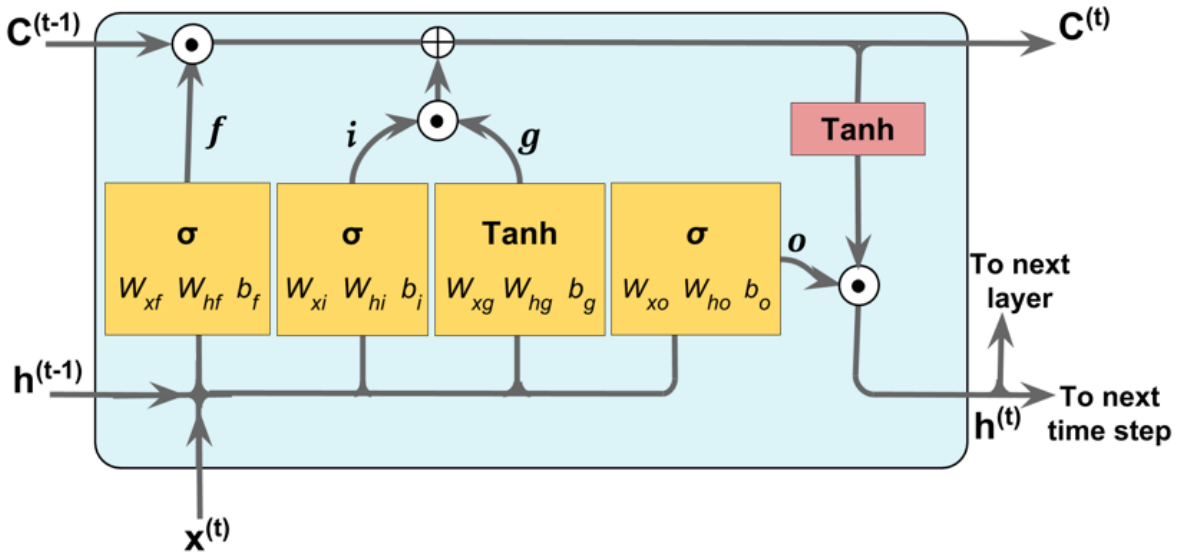
W_{xh} ・・・入力$x^{(t)}$と隠れ層hの間の重み行列
W_{hh} ・・・リカレントエッジに関連付けられた重み行列
W_{xh} ・・・出力層の重み行列



リカレントエッジ
隠れ層の連続する時間刻みの間を情報が流れることで、
ネットワークが過去の事例に関する記憶を持つことが可能となる.

LSTMユニット

↓　現在の入力に重みづけを行う



LSTMの構成要素　メモリセル

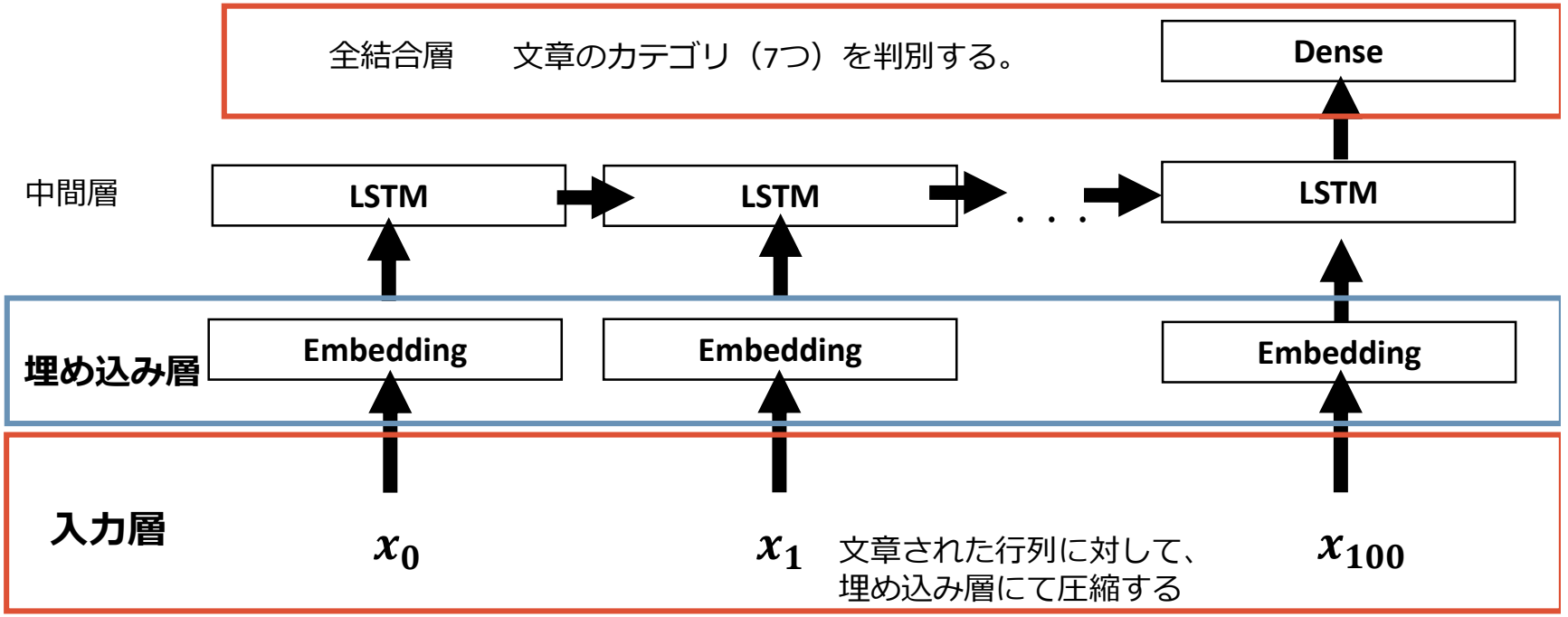
忘却ゲート f
セルの状態をリセットすることが出来る

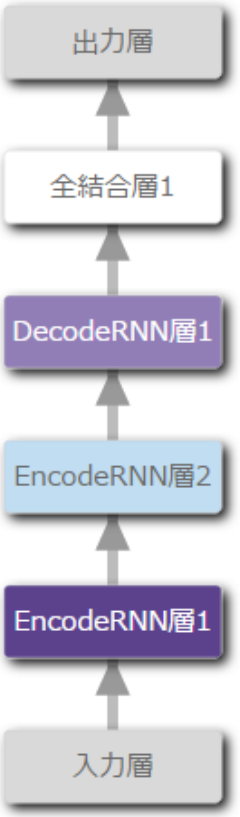
入力ゲート i
入力ノードでセルの状態を更新すること出来る

出力ゲート o
隠れ層のユニットの値の更新方法を決定する

これらの３つから構成される.

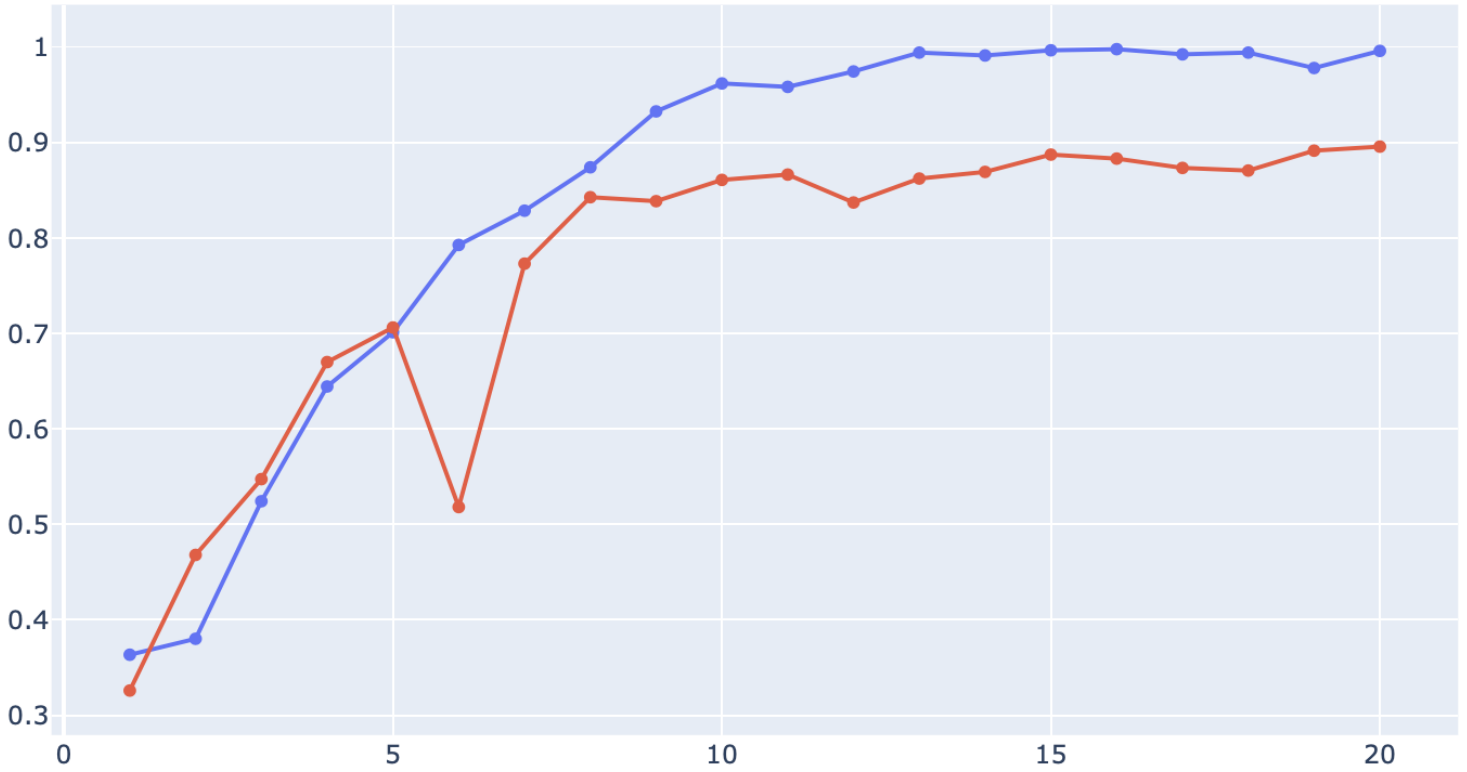
学習の流れ





シーケンス（系列データ）　＝　レビュー

x_0	x_1	x_2	x_3	・	・	・	・	・	・	・	x_{99}	x_{100}

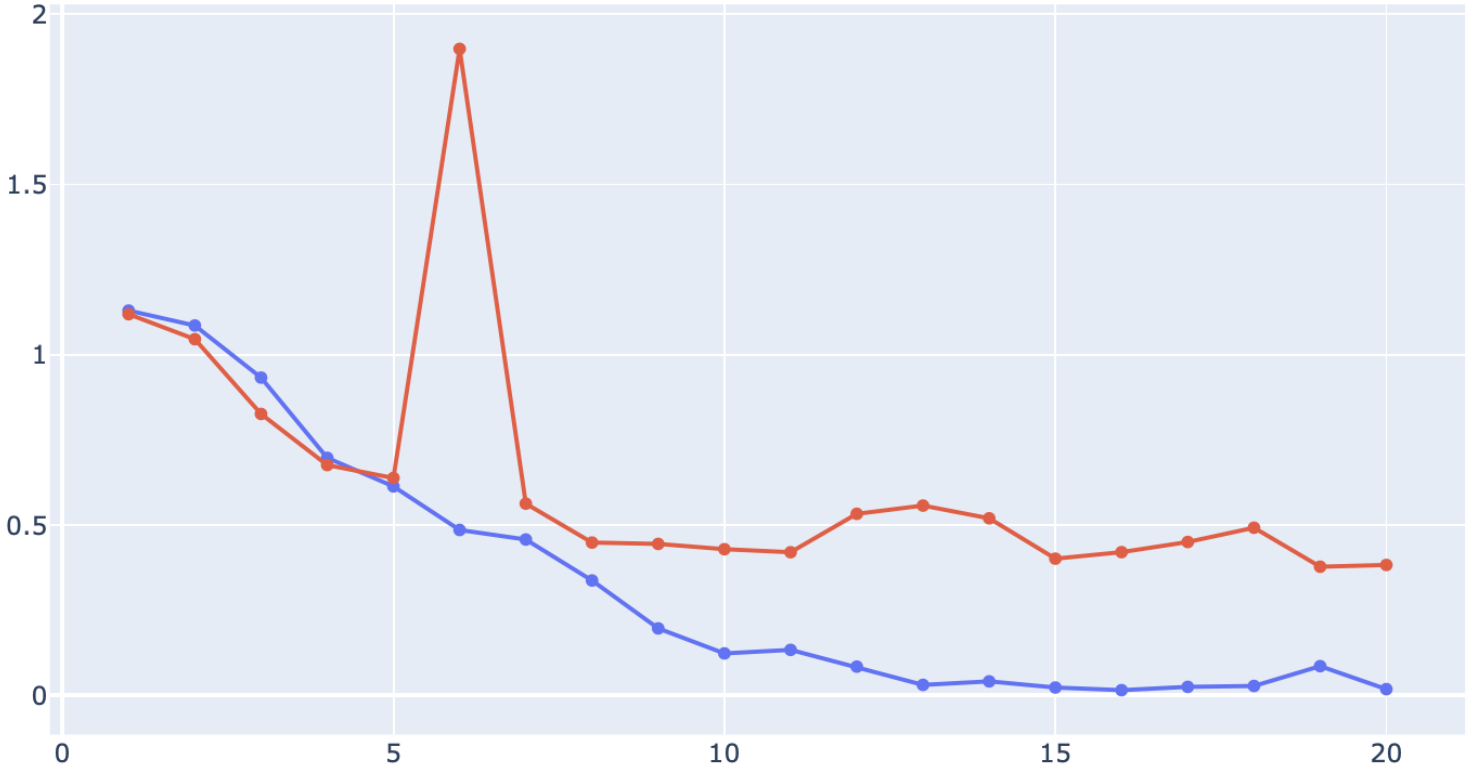


モデル精度の推移

20エポックの学習を行った.

過学習も少なく,
モデル内で学習を行っているとわかる.

3値の分類問題にて,
最終な精度はaccが
94.95% となる.



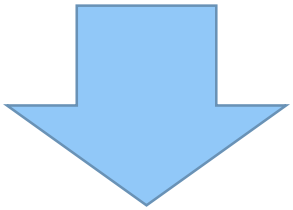
損失関数の推移

損失関数の推移を表す.

全体的に収束していることから, 学習はうまくできていると考えられる.

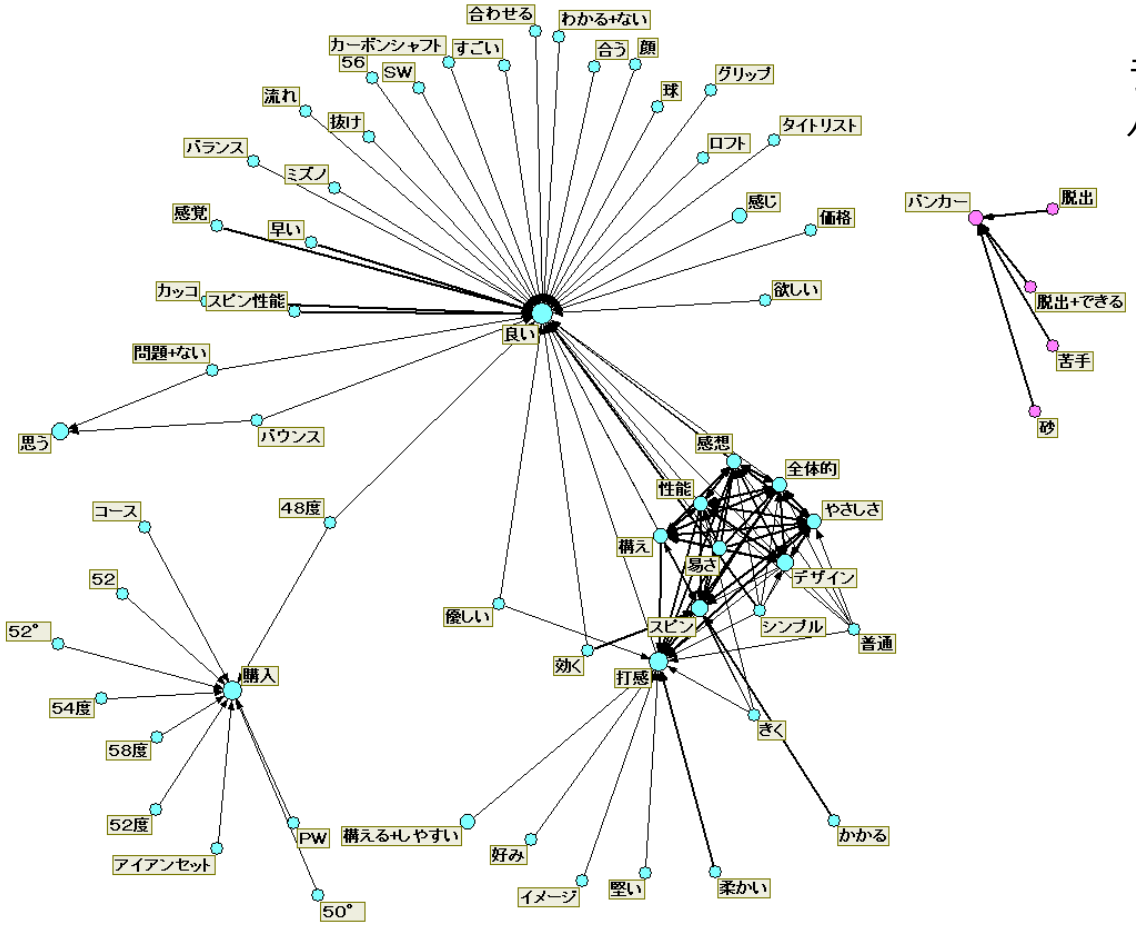
3値の分類問題にて, 最終な精度はlossが17.42%となった.

ここまでの結果より，
単語の並びによってレビュー内容を判別することが可能であると判明.



また，分類モデルでは，かなりの精度で判別できたことからレビュー内にどのような単語が含まれていたのかを明らかにする.

単語ネットワークを構成
・単語が共起されている回数を重みとしたネットワークを作成本分析では，
5回以上共起している単語を選択する.

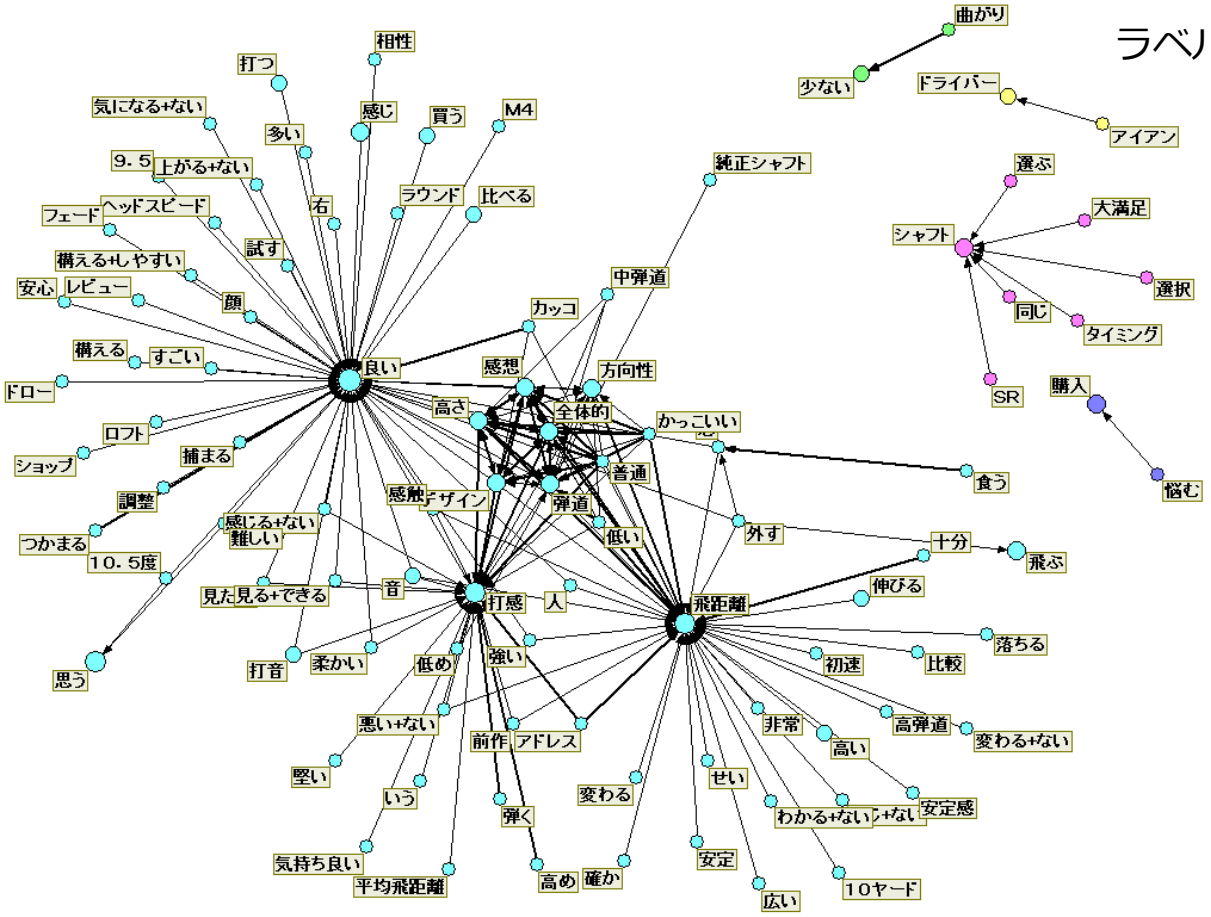


ラベル1（ウェッジ）の言葉ネットワーク
ハブとなっている３単語を選択

良いの共起単語
商品名や”バランス”，”性能”，”グリップ”など，商品が持つパフォーマンスによる単語が集中していることがわかった．

購入の共起単語
数値との共起が見られた．
購入した商品の度数を表す表現が多いことがわかる．

打感の共起単語
”柔らかい“，”硬い”のように，感覚を表すが共起している．

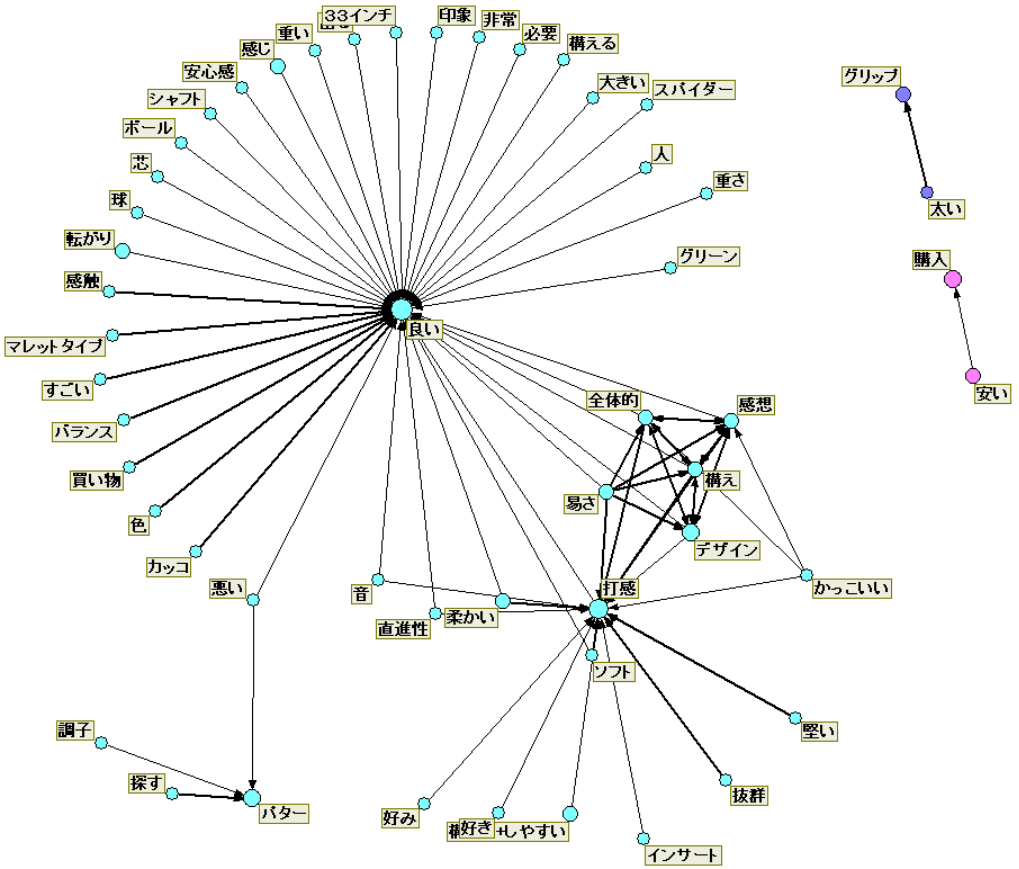


ラベル2（ドライバー）の言葉ネットワーク

良いとの共起表現
“比較”，”試す”などの単語が多く含まれている．
また相性などの単語もあることから，顧客の購買選択基準として，比較した結果で良いものを買うなどの購買パターンが考えられる．

打感との共起単語
“強め”，“音”，“デザイン”など、感覚を表す単語が算出された．

飛距離との共通単語
“安定”，”初速”など，商品が持つ性能に関した単語が選択された．



ラベル３（パター）の言葉ネットワーク

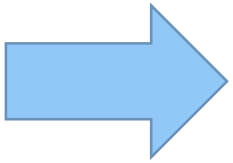
良いとの共起表現
商品のビジュアルを表す単語が多い
“印象”，”カッコ”，”色”など
また，”安定感”や，”必要”，”重さ”といった
単語から，商品自身のスペックを示す
単語が選択された．

打感との共起表現
好み、”ソフト”、”インサート”など

ウェッジのレビュー
性能やバランスといった明確に商品自身のパフォーマンスを表す単語が類推された.
今回で言えば“性能”，などの**商品性能を表す特定の表現**が多く出ていた.

ドライバーのレビュー
比較や相性などの単語から，**他のものとの比較や，併用の可能性**が示唆される.

パターのレビュー
ビジュアルなどの単語や，好みなどの単語から，比較的**商品単位の好み**に関する内容が示唆される.



判別モデルでは，単語自体と
単語の並びの情報を元に学習を行っている.

本分析では,
単語を時系列データとして判別モデルに入れ,
内容から商品のカテゴリラベルを推測した.
また，ネットワーク分析によって、レビューにどのような特徴があるのかを明らかにした.

ネットワークの分析により，カテゴリによって対象としている
内容に違いがあることがわかった.
この内容と利用している単語の違いによって，モデルが判別できたと考えられる.

今後の課題として、
レビューの内容だけでなく，
各ユーザー の商品好みなども考慮し，モデルを構築する.

レビューカテゴリでなく，レイティングデータと合わせて、
どのような表現が良い評価につながっているのかを考察する.

参考文献

[1] S. Hochreiter; J. Schmidhuber (1997). *“Long short-term memory”, Neural Computation* **9** (8): 1735–1780.

[2] Gers, F. A.; Schmidhuber, J. (2001). “LSTM Recurrent Networks Learn simple Context Free and Context Sensitive Languages”

[3]松尾　豊, 石橋　満「語の共起の統計情報に基づく文書 からのキーワード抽出アルゴリズム」人工知能学会誌 17（3）, 2002.