常識から外れた雑談応答の検出
Detection of chat response that is out of common sense

林 卓矢　　荒木 雅弘

京都工芸繊維大学

hayashi@ii.is.kit.ac.jp, araki@kit.ac.jp

1　はじめに

社会規範から外れ，相手を不快にするような，いわゆる常識から外れた発話を非常識発話と呼ぶ．雑談対話システムが非常識発話を出力してしまうと，雑談対話システムを作成している企業や個人の信用を著しく失う危険性がある．

しかし，人手で非常識発話を検出する識別器を構成することは非常にリスクが高い．なぜならば，非常識発話を多く見ることで開発者の心身に悪影響を与える可能性があるからである．そこで本研究では，リスクを低く抑えつつ非常識発話を自動的に検出するための識別器の構成手法を提案する．

2　非常識発話について

社会規範から外れ，相手を不快にするような発話を非常識発話と定義する．非常識発話の例として以下のものが挙げられる．

- 対話相手を意図的に傷つけるもの
- 対話相手を侮辱するもの
- 対話相手の嫌がる話題を含むもの

本研究での識別対象は対話中のシステム発話とする．その発話が，非常識発話であるか非常識発話でないかを識別することが目標である．

対話内容を変数で表すと，以下の通りとなる．S はシステム発話の単語系列，sw_i はユーザ発話内の 1 単語を示す．

$$S = sw_1 sw_2 ... sw_i ... sw_n (n \text{ は } S \text{ の全単語数}) \quad (1)$$

非常識発話であることを検出したいシステム発話は S であり，非常識発話の検出に用いる変数は，システム発話の単語系列 S のみとする．文脈情報を加えて非常識なシステム発話の検出を行いたかったが，問題を簡単にするため本稿では発話からの情報のみに限定して取り組んだ．

非常識発話の識別器を作成するにあたって，非常識発話に対する識別器の識別性能を評価する必要がある．識別器を評価する際，雑談対話システムに於いて，最も重視すべき変数は $Recall$ であるが，全てを非常識破綻であると識別することで $Recall$ が最大となる識別器は構成できてしまう．そこで，本研究では識別器を評価する変数を $F - measure$ とする．

3　関連研究

磯野ら [1] は Web 掲示板から，皮肉および誹謗中傷を自動検出する手法を提案している．皮肉の検出では，皮肉を 8 つに分類し，それぞれに対して有効な手法を用いている．誹謗中傷の検出では，誹謗中傷語リストを作成している．識別対象の発話とその直後，直前の発話に出現する誹謗中傷語句の出現頻度を素性として，SVM(Support Vector Machine) に入力することで識別対象の発話が誹謗中傷であるかを識別している．

松葉ら [2] は学校非公式サイトにおける有害情報の検出実験を行った．有害情報の検出実験では，誹謗中傷の検出のために有害情報を含む書き込み 750 件から，主要素となる単語 216 単語を人手で抽出している．この単語群と，SVM を用いることで有害情報検出を行っている．また，有害情報であるかの判断基準については，文部科学省の有害情報類型化に基づいており，その類型化を基に有害情報であるかどうかを人によって判断している．

前述した 2 つの研究では，単語リストを構築しているが単語の出現回数のみを特徴として用いている．本研究では，単語が出現しているかの情報に加えて，表

記ゆれや類似単語を含めて非常識発話を検出するために Word2Vec[3] を用いる．さらに，Word2Vec を用いることで非常識単語にのみ存在する単語の傾向などが獲得できる可能性がある．

4　提案手法

4.1　非常識単語リストについて

非常識破綻である発話には，単語の出現の仕方に何らかの傾向があると仮定する．本研究ではこの仮定に基づいて，非常識破綻である発話を検出する識別器を構成する．

非常識単語リストの構成方法を図 1 に示す．西原ら [4] は，電子掲示板から文脈を考慮した誹謗中傷文を検出するにあたって，他者を誹謗中傷する可能性の高い 42 単語のリストを作成した．(以下，リスト 1 と表記する．) 本研究では 42 単語だけでは全ての非常識発話を検出するには足りないと考え，Word2Vec を用いてリストを拡張する．

Word2Vec モデルの学習においては，匿名掲示板である「おーぷん 2 ちゃんねる」[1] から取得した人気記事 7959 記事を用いている．人気記事とは，最も近い時点に更新があったものでかつ，50 件以上の投稿があったものである．[2] 取得した記事から，ウインドウサイズを 15，ベクトルのサイズを 200，登録単語の最低出現回数を 20 として Word2Vec のモデルを構築した．Word2Vec のモデルを用いて，リスト 1 内の各単語とコサイン類似度の高い上位 30 単語のリストを作成し，リスト内の単語から人手によって非常識単語であると判断された 362 単語のリストを作成した [3]．(以下，リスト 2 と表記する．) このようにして二つの非常識単語リストを構成した．リストの内容を変数で表すと，以下の通りとなる．D はシステム発話の単語系列，dw_j はユーザ発話内の 1 単語を示す．

$$D = dw_1 dw_2 ... dw_j ... dw_m (m \text{ は } D \text{ の全単語数}) \quad (2)$$

4.2　発話のスコアリング

一つの発話を形態素解析し，単語ごとに前節で構築した Word2Vec のモデルを用いて，ベクトルに変換する．また，非常識単語リスト内の単語も全てベクトルに変換する．Word2Vec 関数は引数の単語全てを，ベクトルに変換し，単語ベクトル系列を返す．

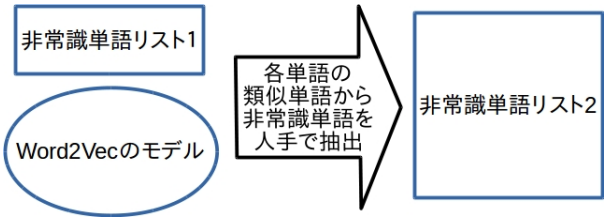

図 1: 非常識単語リストの構成方法

$$\{sv_1, sv_2, ..., sv_i, ..., sv_n\} = Word2Vec(S) \quad (3)$$
$$\{dv_1, dv_2, ..., dv_i, ..., dv_n\} = Word2Vec(D) \quad (4)$$

単語のベクトルを用いて，各発話の非常識度合いのスコアを $Score$ として式 5 に示す．

$$Score = Func2 \quad (Func1(cos(sv_1, dv_1), cos(sv_1, dv_2), ...), \\ Func1 \quad (cos(sv_2, dv_1), cos(sv_2, dv_2), ...), ...) \quad (5)$$

$cos(v1, v2)$ は，$v1$ と $v2$ のコサイン類似度を示す．$Func1$ と $Func2$ は，識別器ごとに違う組み合わせの集約関数を適用する．本研究で用いる組み合わせは次の 4 つである．

- Func1：総和，　Func2：総和
- Func1：最大，　Func2：総和
- Func1：最大，　Func2：最大
- Func1：総和，　Func2：最大

次章では，非常識発話の自動抽出を行うために Func1，Func2，非常識単語リストの組み合わせを変えながら，それぞれの識別器の性能を評価する．

5　実験

5.1　評価用データの収集方法について

評価用データを集めるため，22 名の学生に NTTDOCOMO 雑談対話 API[5] と対話してもらった．実験参加者には，1 対話で 10 回の発話を行ってもらい，システムの発話を除く実験参加者の 3，6，9 発話目には非常識発話をするよう指示した．しかし，実験担当者がどの実験参加者がどのような非常識発話を行ったか知

[1] http://open2ch.net/menu/
[2] 2016 年 11 月 19 日に取得
[3] 非常識発話を見ることを問題と考え，非常識単語を見ることは問題でないと考えた．

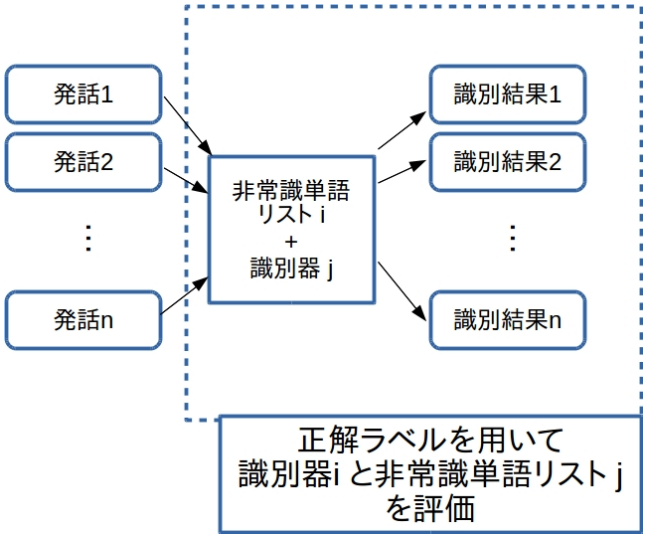

図 2: 識別器の評価方法

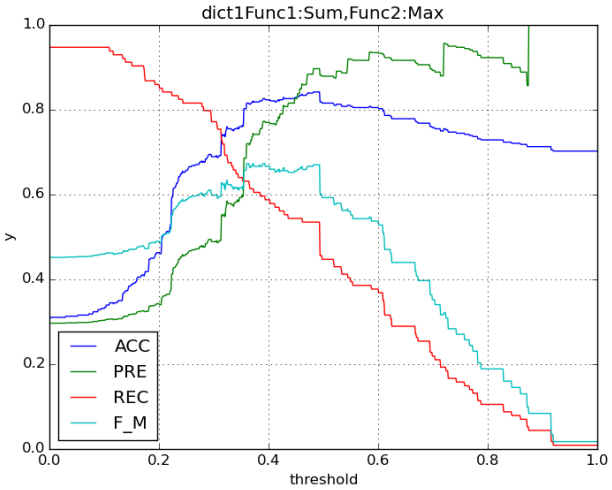

図 4: Func1:Sum,Func2:Max,list1 の性能

ることが出来る状態ならば，収集したデータにバイアスがかかる危険性がある．そこで，事前に実験参加者には対話システムの操作方法を説明しておき，実験担当者がいつ，どの実験参加者が対話を行っているか分からないように対話データを作成した．

5.2 識別手法の評価について

非常識発話を行ってもらった発話を正例，非常識発話以外の発話を負例として評価用データを収集した．収集したデータは，正例 117 発話を含む 390 発話である．

それぞれの識別器，非常識単語リストごとの識別器の性能を評価する．評価の流れを図 2 に示す．性能評価の指標として，ROC 曲線における AUR(Area Under ROC curve) を用いる．

5.3 結果

各識別器，各リスト毎の ROC 曲線を図 3，AUR を表 1 に示す．各図表における list1，list2 はそれぞれリスト 1 とリスト 2 を指す．最も，AUR が高い数値を持つ識別手法は，Func1 が Sum，Func2 が Max を持ち，リスト 1 を用いた場合であった．Func1 が Sum,Func2 が Max の組み合わせの AUR が最も高かった理由として，発話内に一つでも非常識単語が含まれていれば，高いスコアを返し，発話内の他の単語を無視できるためだと考えられる．次に，リスト 1 と リスト 2 のスコアが大きく変わらなかった理由として，リスト 2 内においてコーパス内で出現回数の少ない非常識単語のベクトルが，本来あるべきベクトルと大きく離れて訓練されたため，良い結果に繋がらなかったと考えられる．

また，この識別手法を用いて，正例負例の判断基準となるしきい値を変化させながら *Precision*(PRE) と *Recall*(REC)，*Accuracy*(ACC)，$F-measure$(F-M) の遷移する様子を図 4 に示す．識別対象となる発話に対し，ランダムに識別結果を返す手法をベースラインとして考える．ベースライン手法では，F 値は 0.375 である．本研究では図 4 より，ベースライン手法に比べて高い $F-measure$ を獲得できた．しかし，本研究でテストに用いた評価用データは実験参加者によって作られた質の高いデータであったため良い結果であったと考えられる．本研究ではデータを見ることなく，システムのチューニングを行えることを示した．

6 おわりに

本研究では，データを見ることで研究者の心身に与える悪影響が危険性があるという問題に対し，データを見ることなく，識別器のチューニングを行えることを示した．しかし，1 発話についてのみを対象とした非常識発話検出を行っており，文脈を考慮するなどの複雑な問題に取り組めていない．今後は，複雑な問題に於いても同じようにデータを見ることなく，識別器を構成できることを示すことを目標とする．

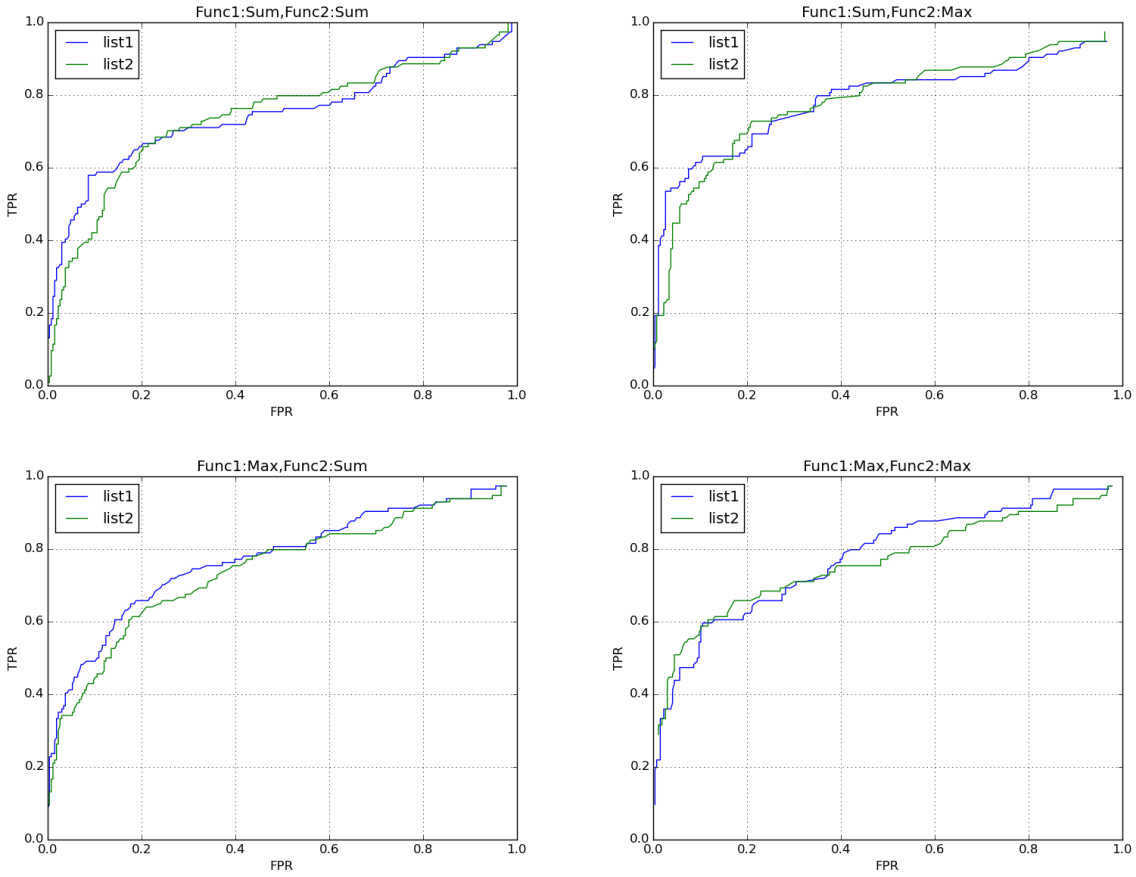

図 3: 各識別手法の ROC 曲線

表 1: 各手法における AUR

AUR	Func1:Sum,Func2:Sum	Func1:Max,Func2:Sum	Func1:Sum,Func2:Max	Func1:Max,Func2:Max
リスト 1	0.747	0.773	0.786	0.777
リスト 2	0.742	0.744	0.784	0.764

参考文献

[1] 磯野史弥, 松吉俊, 福本文代. Web 掲示板における皮肉の分類および自動検出. 研究報告自然言語処理（NL）, Vol. 2013, No. 7, pp. 1–8, sep 2013.

[2] 松葉達明, 桝井文人, 河合敦夫, 井須尚紀. 学校非公式サイトにおける有害情報検出. 信学技報, Vol. 109, pp. 93–98, 2009.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013.

[4] 西原陽子, 岩佐一樹, 福本淳一. 電子掲示板からの文脈を考慮した誹謗中傷コメントの抽出. 人工知能学会全国大会論文集, Vol. 28, pp. 1–4, 2014.

[5] 大西可奈子, 吉村健. コンピュータとの自然な会話を実現する雑談対話技術. NTT DoCoMo テクニカル・ジャーナル, Vol. 21, No. 4, pp. 17–21, jan 2014.