# 会話ログ分析による
# ユーザーごとの関心に応じた話題提案手法

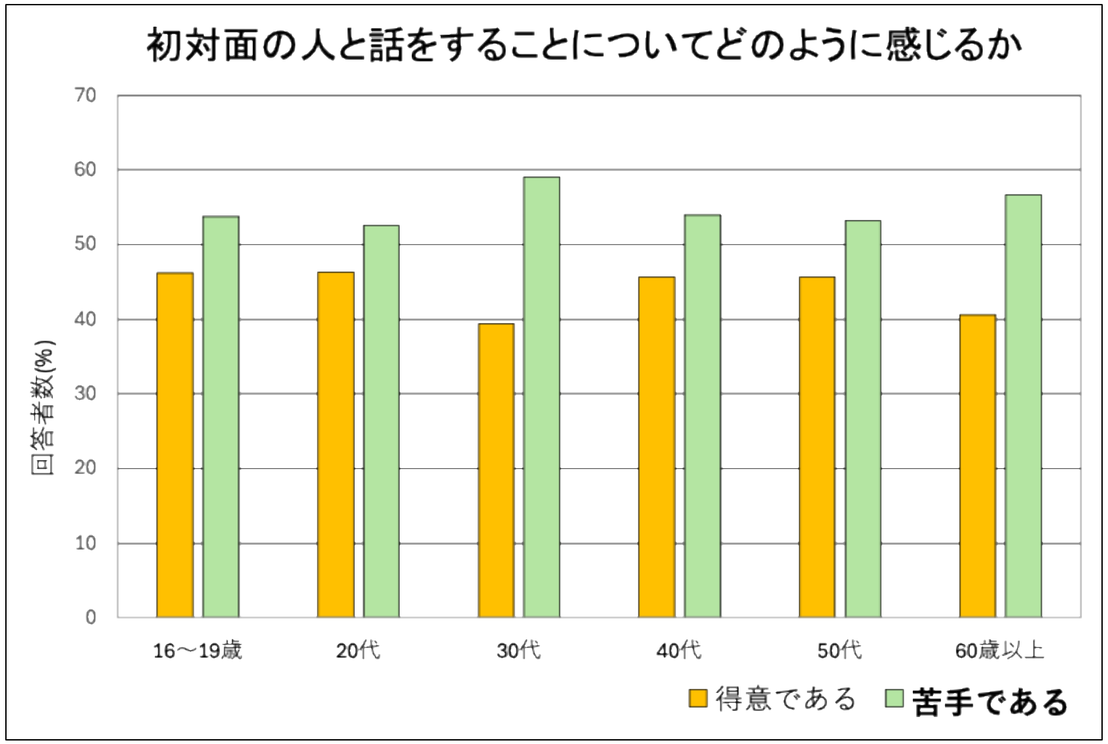
武蔵野大学院 工学研究科 数理工学専攻1年佐々木研究室

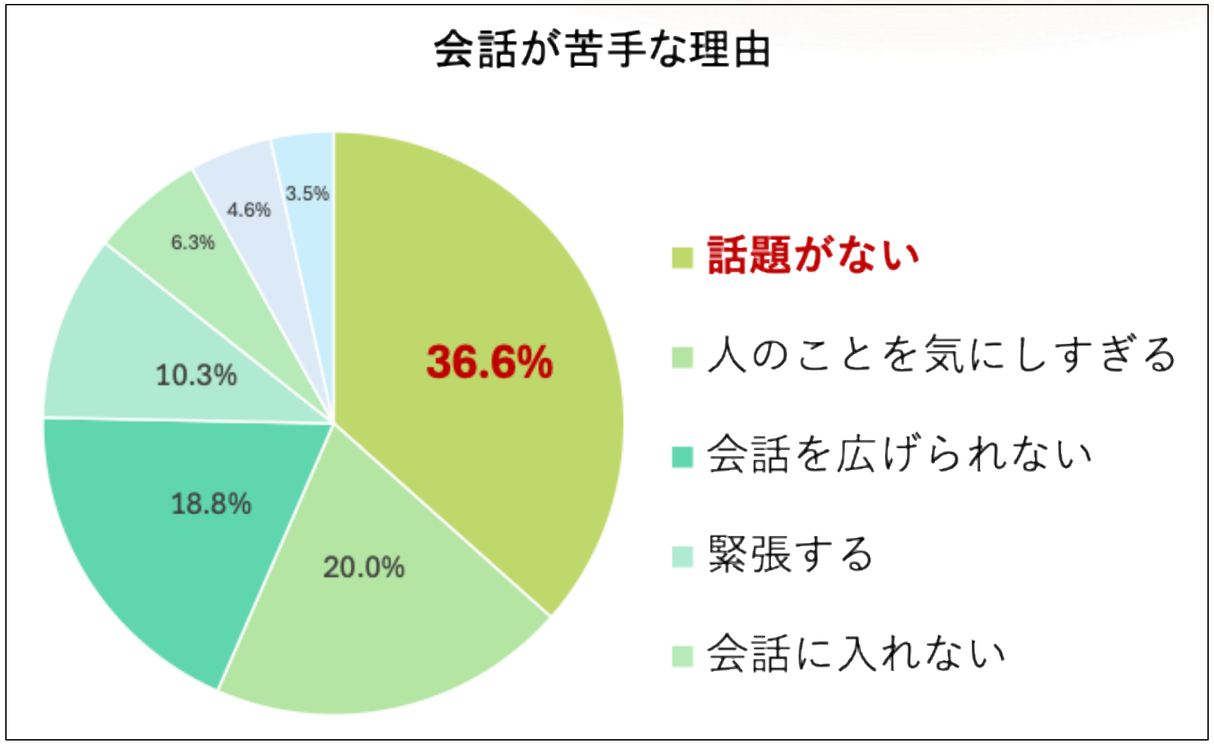
白川桃子

指導教員　佐々木多希子

# 初対面の人って何を話せばいいの？

# 01　研究背景




どの世代の人も初対面の人との**会話に苦手意識**があり**何を話せばいいか分からない**のが原因

# 02　本研究の目的

あなたの雑談を分析

↓

興味のあるトピックを抽出

↓

**初対面の人と話すときに**
**あなたにとって最適な話題を提案！**

# 02　本研究の目的

**目的 - 1**　**ユーザーの関心に基づいた**話題提供の実現

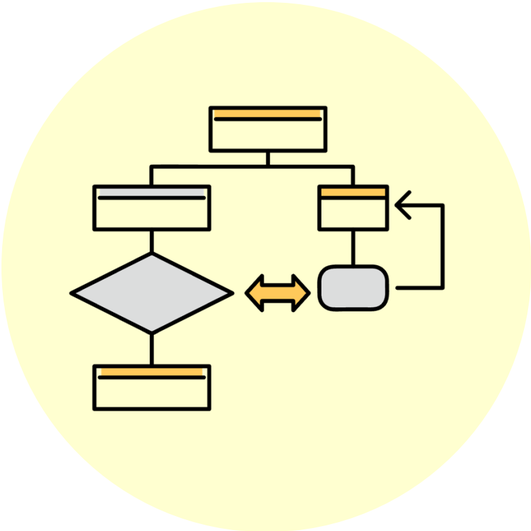
**目的 - 2**　ユーザーの**興味関心を定量的に**評価する

**目的 - 3**　話題提供をすることで初対面時の**会話の盛り上げる**

# 03　自然言語処理とは

**自然言語処理(NLP)**：人々が日常的に用いる言語(=自然言語)を分析対象とし，
その意味を正しく解析できるようにする技術

- **自然言語理解(NLU)**：機械が文章の作りや意味を解析することで文章を判別
- **自然言語生成(NLG)**：入力されたデータに基づき機械が応答する文章を生成する



ルールベース

統計ベース

機械学習ベース

# 04　本研究で活用するデータ

名古屋大学の会話データを活用
(不必要な情報を削除し，DataFrameでデータ整理)

@データn番（会話時間）
@収集年月日：xxxx年xx月xx日
@場所：会話場所
@参加者ナンバー：性別，年齢，出身，所在
.
.
@参加者の関係：関係性
参加者ナンバー:会話内容
.
.

%com:　補足情報
()　相槌
<笑いor間>　自分の笑い，ある程度の沈黙

前処理

| 収集年月日 | 場所 | 追加 | 時間 | 参加者 | 会話内容 |
|---|---|---|---|---|---|
| 2001-10-16 | ファミリーレストラン | None | 35 | M023/F107/F023/F128 | F107/＊＊＊の町というのはちいちゃくって、城壁がこう町全体をぐるっと回ってて、それが城壁... |
| 2001-10-16 | ファミリーレストランガスト | None | 60 | F107/F023/F128 | F107/今度はーイギリスにもアメリカと同様のテロが起こるだろうって言ったんだってよ。 F1... |
| 2001-10-23 | 車中（某大から所属大学への帰り道。運転者F033） | None | 43 | F056/F033 | F033/倒れちゃう。 F056/いきなり倒れた。 F033/どうしよう。あっ、この間に。 ... |
| 2001-10-23 | 車中（知立駅より西尾市まで。運転者M018） | None | 35 | M018/F128 | F128/いや、別にいいよ。ローソンでいいやろ。ちょっと倒していい、これ。どうよ、調子は。 ... |
| 2001-10-23 | M023の自宅 | None | 55 | F116/M026/M023/F128 | F128/来てたときによく貸してもらったやつだ。 M023/そう、そんな感じのとこ。 F12... |

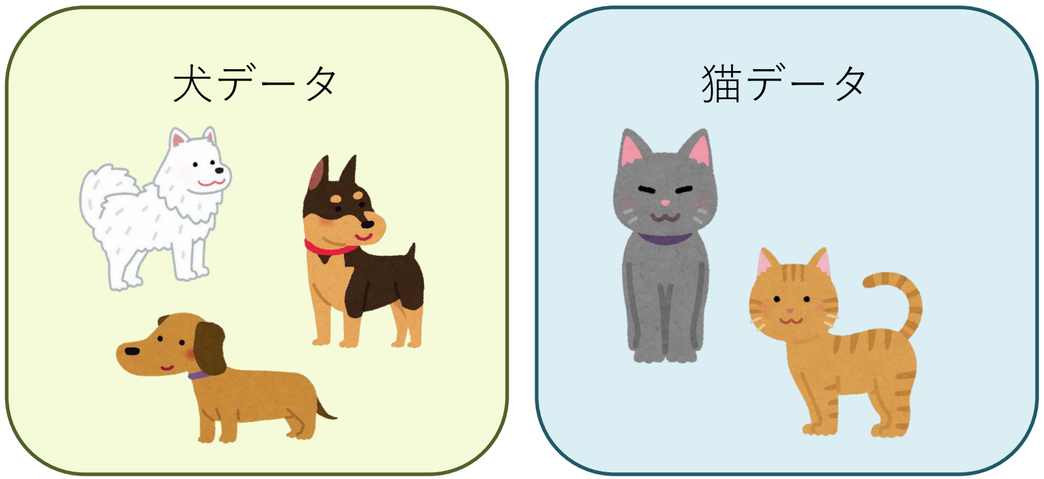
会話1つずつにラベル群から**最もマッチするラベル**を付与したい
会話データを相当量学習する必要があるが，**会話が129個**しかない

**trainデータとtestデータが一致していなくても分類可能な手法を活用**

# 05　ゼロショットラーニング – 概要

**trainデータが存在しないもの**に対して
事前に特徴を学習する形で識別することが可能

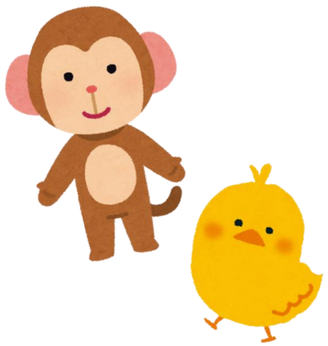
【trainデータ】




【testデータ】

**一般的な学習方法**の場合…
trainデータと
**同じ項目**について予測

**Zero-shot-learning**…
tainデータとは違う
**新たな項目**について予測



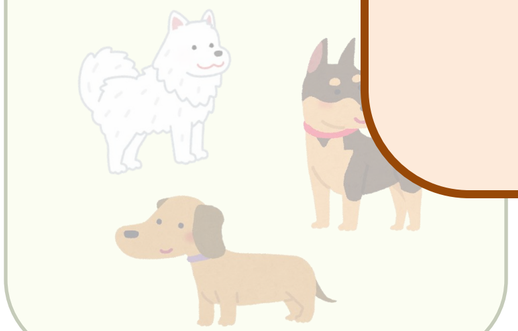
# 05 ゼロショットラーニング – 概要

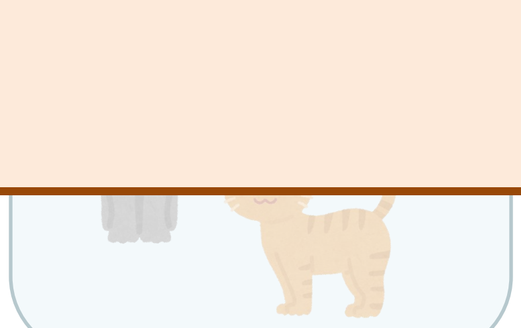
trainデータが存在しないものに対して

【 解決へのアプローチ方法 】

**ラベルの代わりに文章を使う**

犬データ





**Zero-shot-learning…**
tainデータとは違う
**新たな項目**について予測

# 05　ゼロショットラーニング – 概要

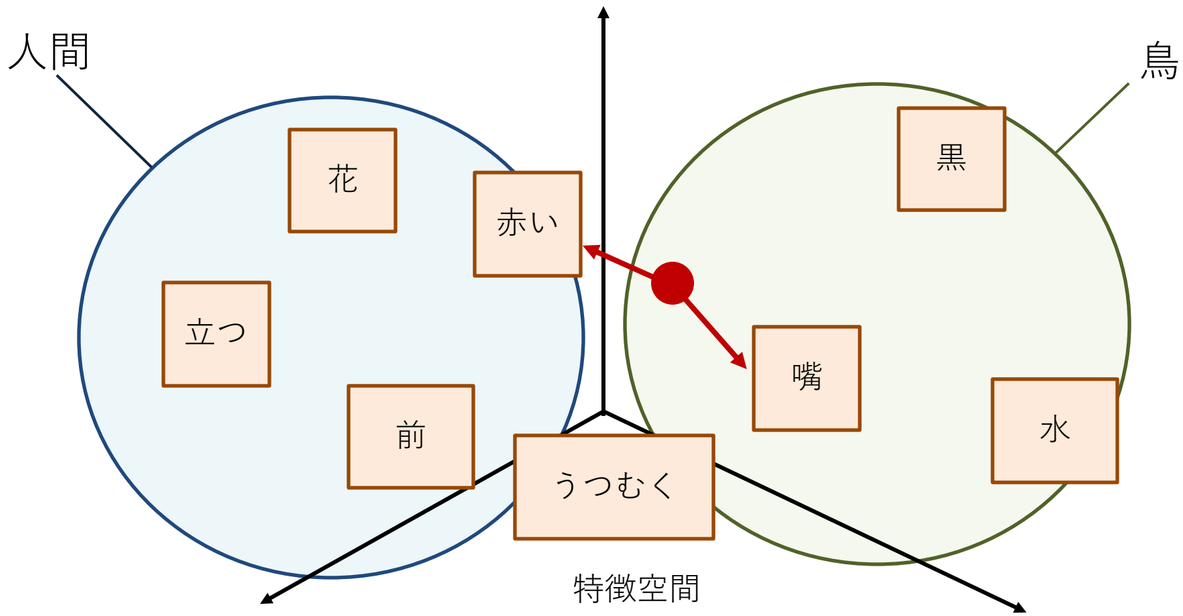
trainデータ

赤いうちわを持って
花の前に立つ女性

黄色い花の前で
うつむく女性

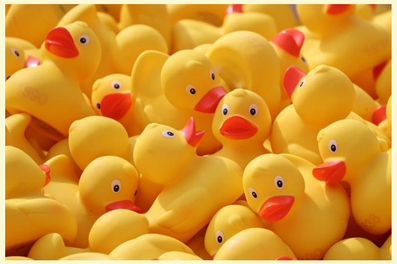
赤い羽と黒い目を
持った鳥がうつむく

水の上で遊ぶ黒い鳥




**細かい情報**まで特徴空間に表現可能

testデータ



推論データに近い特徴を
使い文章を出力

**赤い嘴の鳥**

# 05 ゼロショットラーニング – 問題の定式化

## ◯活用する2つのクラス

$S$：学習用の可視クラス　（本研究ではtrainデータの日本語SNLIデータ．548,014ペア．）
$U$：予測用の不可視クラス（本研究ではtestデータの名大会話コーパス，noteのカテゴリ）

$$S = \{c_i^s \mid i = 1, \ldots, N_s\}$$
$$U = \{c_i^u \mid i = 1, \ldots, N_u\}$$
$$S \cap U = \emptyset$$

## ◯各クラスに含まれる集合

$S$

$D^{tr}$ ：可視クラスのラベル付きインスタンスの集合

$$D^{tr} = \{(x_i^{tr}, y_i^{tr}) \in X \times S\}_{i=1}^{N_{tr}}$$

$X$：特徴空間
$(x_i^{tr}, y_i^{tr})$：各ラベル付きインスタンス
$x_i^{tr}$：$X$内のインスタンス
$y_i^{tr}$： $x_i^{tr}$に対応するラベル

$U$

$X^{te}$：テストデータのインスタンスの集合

$$X^{te} = \{x_i^{te} \in X\}_{i=1}^{N_{te}}$$

$x_i^{te}$：$X$内のテストインスタンス

$Y^{te}$：$X^{te}$に対応する予測られたラベルのクラス

$$Y^{te} = \{y_i^{te} \in U\}_{i=1}^{N_{te}}$$

# 05 ゼロショットラーニング – マッチングモデルの構築

$(x_i^{tr}, y_i^{tr})$の例：　（いちごを食べます，　いちごは食品だ）

↓

BERTエンコーダーへの入力形式に変換
[CLS] いちごを食べます [SEP] いちごは食品だ [SEP]

**「いちごを食べます」という文に対して「いちごは食品だ」は同じような意味を持つか**

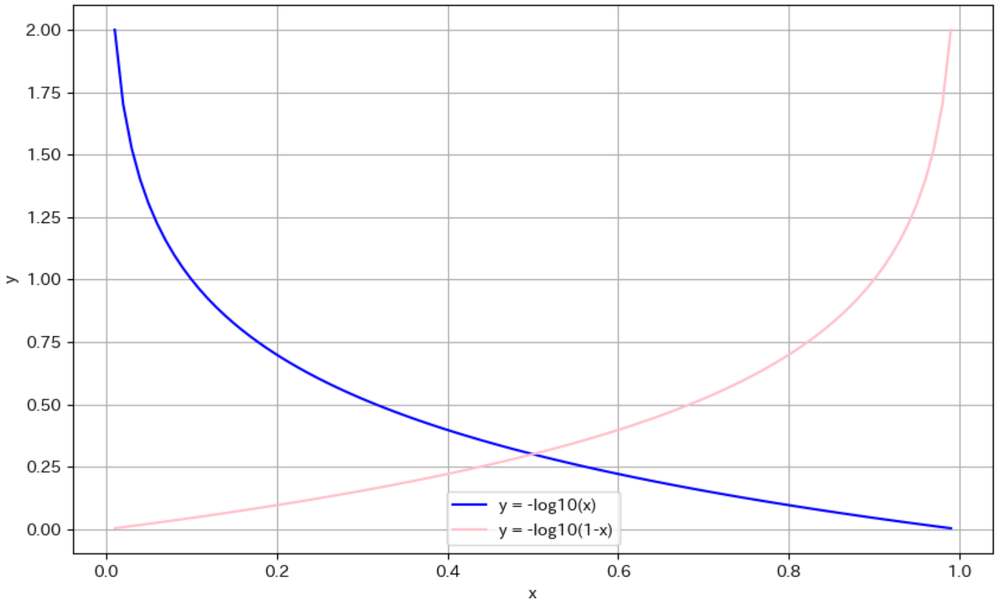
↑こんな感じの操作を$(x_i^{tr}, y_i^{tr})$を変えながら繰り返していく

以下のように線形層(1)を重ね，損失(2)を計算

$$p_{x,y'} = \sigma(W^T c_{x,y'} + b) \qquad (1)$$

$$L = \begin{cases} -\log(p_{x,y'}) & y' = y \\ -\log(1 - p_{x,y'}) & y' \neq y \end{cases} \qquad (2)$$

$W, b$：線形層のパラメーター($W \in R^H, b \in R$)
$c_{x,y'}$：各文章に対応する隠れベクトル($H$次元)
$p_{x,y'}$：$x$と$y'$のマッチングスコア
$\sigma(\cdot)$：シグモイド関数




**$L$が小さくなるようにモデルを構築していく**

# 05 ゼロショットラーニング – 予測

**会話の各文章と「この文章は{noteカテゴリ}に関する会話だ」は似た意味を持つか？**
テキスト$x$とのマッチングスコアが最大となるようなラベル$y^*$を擬似ラベルとする.

| 会話内容 | noteカテゴリ | マッチングスコア | 現時点の$y^*$ |
|---|---|---|---|
| 最近は、女性の方が旅行に行くんじゃないの？ | ショッピング | 0.58 | |
| | **旅行・おでかけ** | **0.79** | ☑ |
| | IT | 0.15 | |
| | 読書 | 0.23 | |

**これらの過程を通して各会話データ(testデータ)に最適なラベルを付与できる！**

# 06 興味関心の定量化

## 記号設定

$N$: 会話データの総数
$P$: 参加者の集合
$L$: ラベルの集合
$S_{i,j}$: i番目の会話での参加者jの発言量
$V_i$: i番目の会話とラベルの一致率
$l_i$: i番目の会話ラベル

## 定量化の流れ

| | アウトドア | ビジネス | まなび | キャリア | 社会 | デザイン | テクノロジー | フード | 読書 | IT | カルチャー | 旅行・おでかけ | ライフスタイル | 経済・投資 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F001 | 0 | 0 | 0 | 23.3335268 | 0 | 0 | 0 | 0 | 0 | 18.9731593 | 6.83649041 | 0 | 0 | 0 |
| F002 | 41.5855285 | 0 | 0 | 14.8541104 | 0 | 0 | 0 | 0 | 25.2971111 | 0 | 0 | 0 | 0 | 0 |
| F003 | 0 | 0 | 41.5111609 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F004 | 33.3821428 | 12.5257818 | 32.908264 | 27.2161248 | 0 | 44.8695885 | 21.5522848 | 0 | 22.8652407 | 18.880105 | 10.4535792 | 0 | 0 | 0 |
| F005 | 0 | 0 | 0 | 0 | 0 | 0 | 21.7936109 | 18.0423357 | 31.2462903 | 0 | 0 | 0 | 0 | 0 |
| F006 | 0 | 0 | 0 | 9.55824153 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F007 | 21.4532295 | 0 | 20.5807441 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F008 | 18.7853362 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F009 | 0 | 0 | 0 | 30.58212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F010 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12.9437646 | 0 | 0 | 0 | 0 | 0 | 0 |
| F011 | 26.2197035 | 0 | 0 | 36.4495779 | 22.7181314 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F013 | 0 | 0 | 0 | 0 | 0 | 0 | 18.7434972 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F014 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4.32142051 | 0 | 0 | 0 | 0 | 0 |
| F015 | 0 | 0 | 0 | 0 | 0 | 6.23773392 | 31.321449 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F016 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26.1424913 | 0 | 0 |
| F017 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18.7061806 | 0 | 0 | 0 | 0 | 0 |
| F018 | 0 | 0 | 0 | 11.3300131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F019 | 0 | 0 | 0 | 0 | 0 | 27.7352345 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F020 | 36.7794531 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 44.0447593 | 0 | 0 | 0 | 0 | 0 |

### 各会話における関心度

$$Score_{i,j} = \begin{cases} V_i \cdot S_{i,j} & l_i = l \\ 0 & l_i \neq l \end{cases}$$

### 各トピックに対する関心度

$$Interest_{i,j} = \begin{cases} \dfrac{\sum_{i=1}^{N} Score_{i,j}}{n_l} & n_l > 0 \\ 0 & n_l = 0 \end{cases}$$

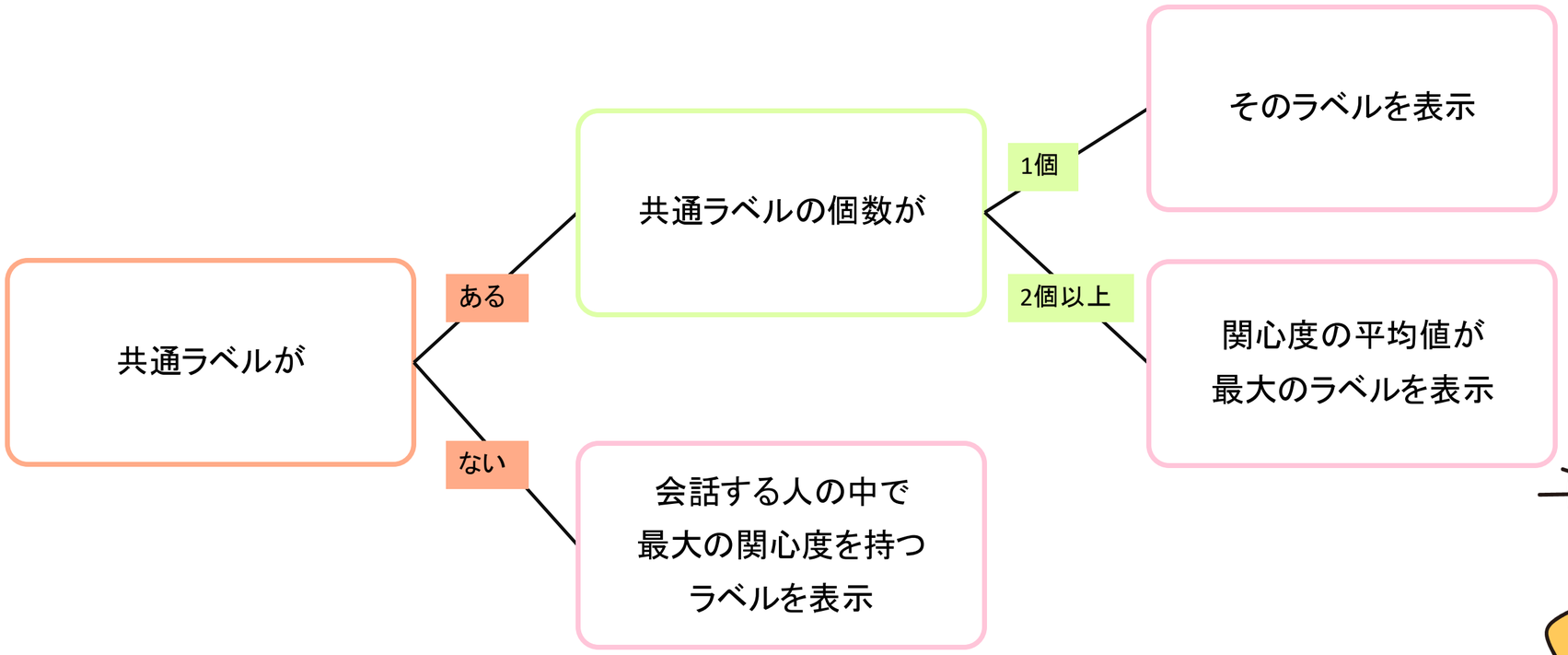
$n_l = \sum_{i=1}^{N} 1_{\{l_i=l\}}$ : 参加者がラベル $l$に関与した会話の総数
$1_{\{l_i=l\}}$: $l_i = l$の場合は1，そうでない場合は0を返す

### 結果の行列式化

$$I = \begin{bmatrix} Interest_{1,1} & \cdots & Interest_{1,|L|} \\ \vdots & \ddots & \vdots \\ Interest_{|p|,1} & \cdots & Interest_{|P|,|L|} \end{bmatrix}$$

# 07　話題とマッチングスコアの提示

マッチングスコアに基づいて各ユーザーの興味関心を定量化する



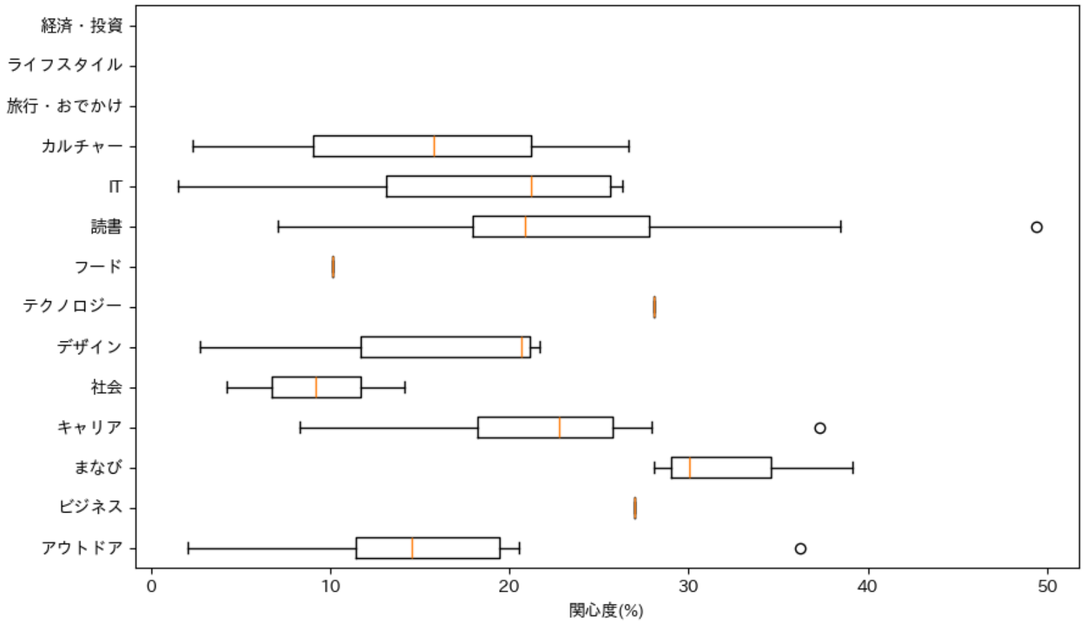

テクノロジー　60%
エンタメ　30%

テクノロジー
興味関心スコア平均：70%

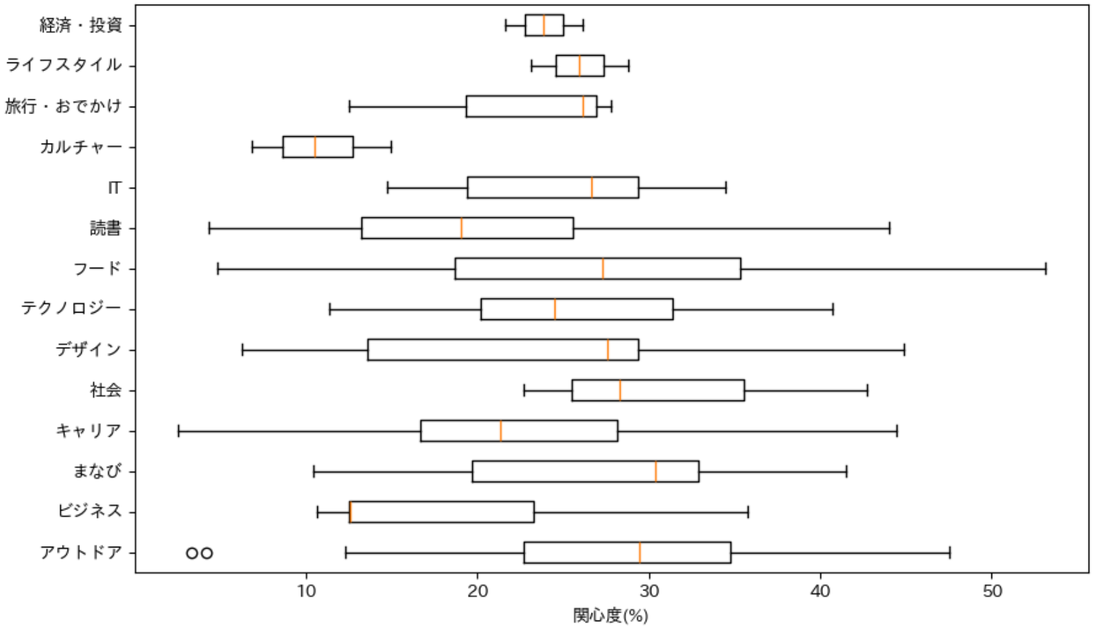
テクノロジー　80%
アウトドア　40%

# 08　評価 –testデータのラベル分布

男性




女性




・女性の方が男性よりばらつきが大きい
　　→**女性の方がトピックについての知識が浅くともその会話に参加する**可能性を示唆

・男性はデータ分布が関心度の低い方に集中しているが，上振れした外れ値が女性よりも多い
　　→これは**男性の方が1つの分野について極端に詳しいという専門性**の高さを示す結果となった.

# 08 評価 – 実験の設定

**被験者**　計11名(男性7名，女性4名)

**実験手順**
(イ)オンライン通話サービスoviceに被験者が入室.
(ロ)文字起こし機能を使いながら10分程度会話をする.
※被験者が初対面の場合はラベルを提示する.
(ハ)文字起こしデータをcsv形式で出力する.
(ニ)ラベル候補を示し，(ロ)の会話に適したラベルを選択.
※複数選択可能

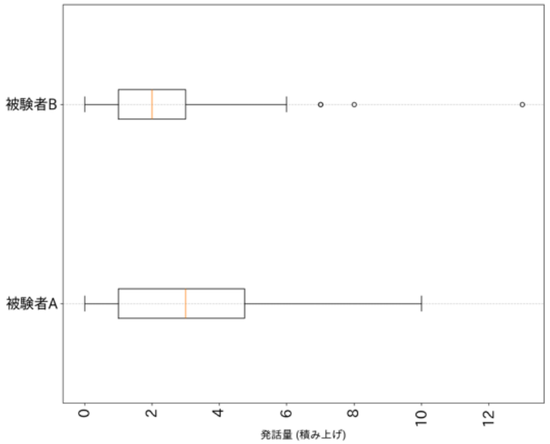
**被験者組み合わせ**　知り合いの会話：7会話
**初対面の人との会話 【共通ラベルあり】：3会話**
**初対面の人との会話 【共通ラベルなし】：1会話**
※各会話の参加者は被験者中から2,3人抽出する

| 被験者 | 年齢 | 性別 | 出身地 |
|---|---|---|---|
| 被験者A | 20代前半 | 女性 | 埼玉県 |
| 被験者B | 20代前半 | 女性 | 東京都 |
| 被験者C | 50代前半 | 男性 | 奈良 |
| 被験者D | 20代前半 | 男性 | 千葉 |
| 被験者E | 60代前半 | 男性 | 新潟 |

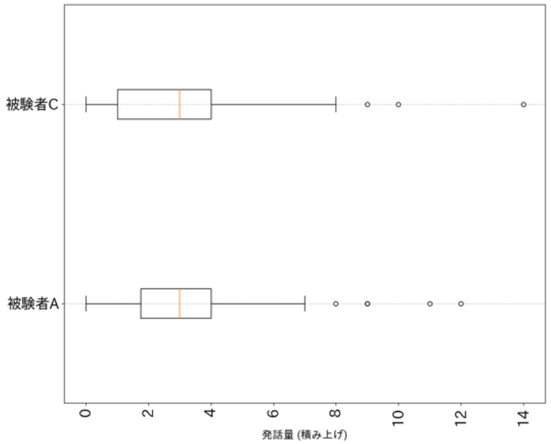
実験実施例

# 08　評価 – 初対面会話時の発話量

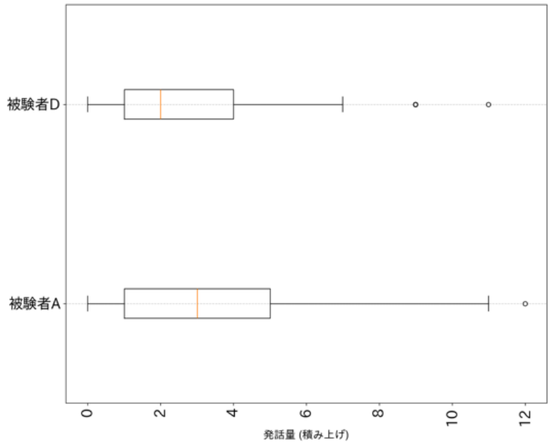
話題提供なし



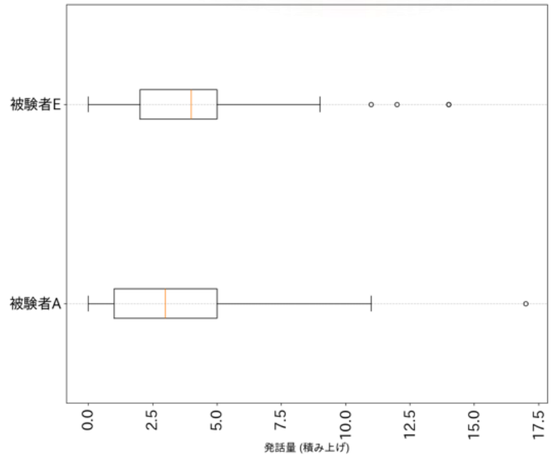

話題提供あり




話題提供あり




話題提供あり




・第一四分位範囲と第三四分位範囲について大差はない
　→どの会話においても**発話量の傾向にも大差がない**

・外れ値について
下振れた外れ値：なし
上振れた外れ値：被験者Aに着目すると話題提供がある会話全部で発生
　→話題がないと会話慎重になっており極端に長い発言が発生しなかったのでは

　→**話題提供がある方が自分の意見や経験談を生き生きと発言できる**可能性が示唆

# 08　評価 – 会話中の感情推移

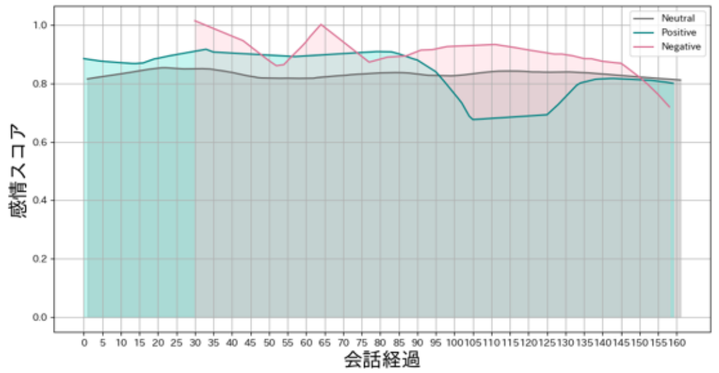
POSITIVE，NEGATIVE，NEUTRALの3感情を分析
(可視化する上でstatusmodelsを用いて平滑化を実施)

ニュートラル
　スコアの値が0.8~1.0で安定

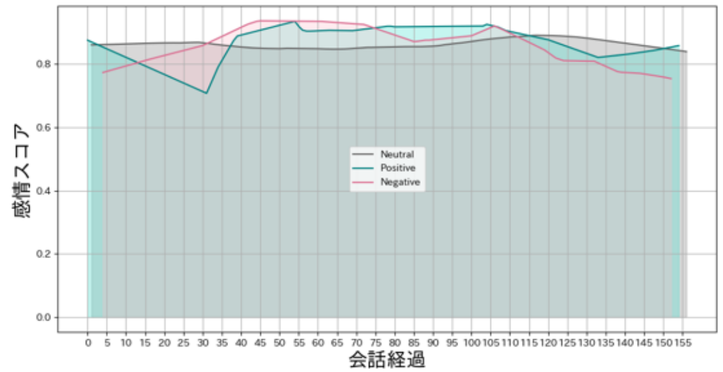
ポジティブ，ネガティブ
　ニュートラルよりも値の変化が大きい
「使ったこと**ない**から気になります」
　→本質的にネガティブでない可能性もあり

話題提供がない時はポジティブ，ネガティブの
面積が**約15%**狭い

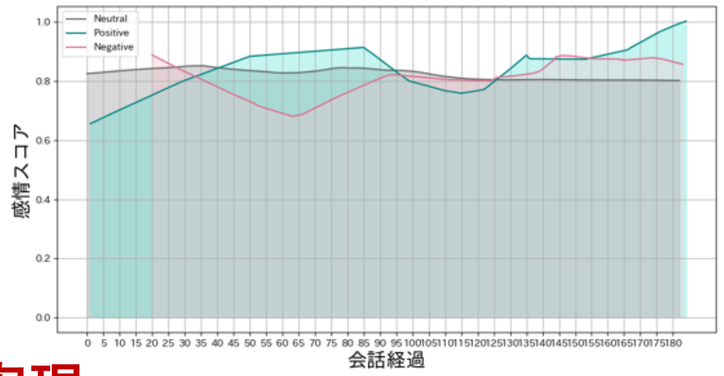
話題提供なし



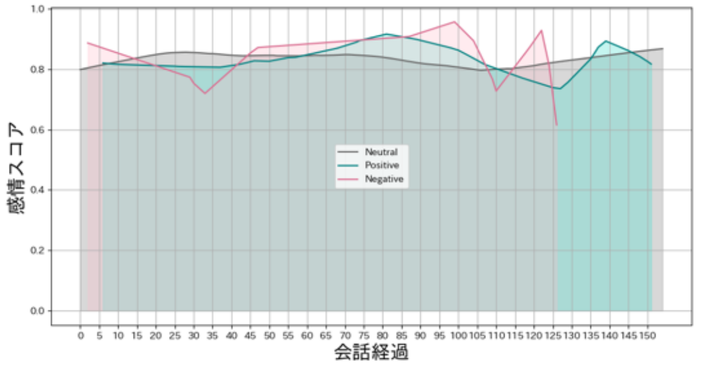

話題提供あり




話題提供あり




話題提供あり




**話題提供がある方が感情豊かに会話が実現**
**　→初対面のうちからより印象に残るインパクトのある会話が実現**

# 09　目標達成の判定

**目的 - 1**

**パーソナライズ**された話題提供の実現

**目的 - 2**

ユーザーの**興味関心を定量的に**評価する

**目的 - 3**

話題提供をすることで初対面時の**会話の盛り上げる**

# 09　目標達成の判定

**目的 - 1**

**パーソナライズ**された話題提供の実現
→会話ログを分析し個々人の興味を抽出・話題選定

**目的 - 2**

ユーザーの**興味関心を定量的に**評価する
→会話のラベルごとに関心度スコアを算出

**目的 - 3**

話題提供をすることで初対面時の**会話の盛り上げる**
→**生き生きと発言し感情豊かな会話**を実現

# 10　今後の展望

| | |
|---|---|
| **展望 - 1** | **会話の盛り上がり評価関数**の作成 |
| **目的 - 2** | 会話を文字データにする**Speech to Text**の実装 |
| **目的 - 3** | 本研究の**サービス化** |

# 11 参考文献


[1]株式会社R&G”会話が苦手なランキング！克服方法も解決”(2024,11,24)
https://r-andg.jp/blog/4733
[2]合同会社Techtale “秒で話題をご提供トークテーマガチャ!!”
https://talkgacha.com/
[3] 小町守(2024) “自然言語処理の教科書” 株式会社技術評論社
[4]白井清昭(2006)” 自然言語処理論Ⅰ 7. 形態素解析(日本語の単語分割)”
[5] 藤村逸子，大曽美恵，大島ディヴィッド義和(2011)
“会話コーパスの構築によるコミュニケーション研究”
藤村逸子、滝沢直宏編『言語研究の技法：データの収集と分析』p.43-72、ひつじ書房
[6]PythonDocs”re --- 正規表現操作”Python3.13.1Document
https://docs.python.org/ja/3.13/library/re.html
[7] Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Xiaoxiao Xu, Suhang Zheng, Feng Wang, Jingmin Chen, Jun Zhang, Huajun Chen” Zero-shot Text Classification via Reinforced Self-training” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3014–3024 July 5 - 10, 2020. c2020 Association for Computational Linguistics
[8]京都大学大学院情報学研究科知能情報学コース言語メディア分野
”日本語SNLI(JSNLI) データセット”(2020,07,15)
[9]note株式会社“note”
https://note.com/
[10]国立国語研究所“現代話し言葉UniDic”UniDic(2023,03,24)
https://clrd.ninjal.ac.jp/unidic/download.html#unidic_csj
[11]HuggingaFace“SentenceTransformers Documentation”
https://sbert.net/
[12] 京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクト
“MeCab: Yet Another Part-of-Speech and Morphological Analyzer”
(2013,02,18)
[13]oVice株式会社”ovice”
https://www.ovice.com/ja