FastText を用いた悩み文へのタグ付け

Concern Tagging using FastText

勝田 暁子 Satoko Katsuta　西村 有紗 Nishimura Arisa　植木 一也* Kazuya Ueki

明星大学
Meisei University

Abstract: We describe one of the functions of building a chatbot system for counseling purposes. The purpose of this study is to automatically estimate what kind of problems counseling users have from their descriptions of their own problems. We trained documents collected from Internet consultation sites using FastText, linked the information on the type of problems to the documents, and used it to improve the accuracy of the chatbot system's responses.

1　はじめに

現在心理的に苦痛を抱えている人が，将来的に専門医にかかるきっかけを作る目的で，人工知能による対話形式のカウンセリングシステムを作成する．これを構成する機能として，大別して画像生成，表情認識，自然言語処理，エージェントの 4 機能が存在し，ここでは自然言語処理の機能の 1 つである悩み文へのタグ付けについて述べる．

2　全体のシステム

本論文は，先に述べたとおり人工知能による対話形式のカウンセリングシステムを構成する機能の 1 つについて述べるものである．このシステムは，様々な要因によって心理的苦痛を抱える人が，将来的に専門医に掛かるきっかけを作ることを目標として，複数人の学生によって現在開発しているものである．

カウンセリングは 2 部構成となっており，前半では自由形式の対話を行うことでシステム利用者の緊張をほぐすことを目的とする．後半ではチェックリストに沿った問診を行うことを通じ，利用者が抱える悩み，ストレスの傾向及びその程度を記録する．このチェックリストは，既存のいくつかのうつ病等の診断を行うチェックリストを参考に作成したものである．最終的にカウンセリングを通じて得られた情報を元に，利用者の現在の心理状態を分析した結果のフィードバックを行う．また，これらのカウンセリングを行うに当たり，人工知能と利用者間のコミュニケーションは文字ベースでのチャット形式で行う．

*連絡先：明星大学 情報学部 情報学科
〒 191-8506 東京都日野市程久保 2-1-1
E-mail: kazuya.ueki@meisei-u.ac.jp

これを実現する機能として，大別して画像生成，表情認識，自然言語処理，そしてそれらの統括を行うエージェントの 4 つの機能が存在する．

画像生成機能では，カウンセリング全体を通じて利用者に対して表示する，対話を行う人工知能そのもののキャラクターイメージを生成する．これは細かなニュアンスを表現することで，利用者に対し安心感や利用そのものの心理的ハードルを下げてもらうことが目的である．

表情認識機能では，カウンセリング中の利用者の顔情報を収集し，カウンセリング全体を通じての緊張度や，人工知能が何の質問をした際にどのような反応をしたか等の自然言語処理と紐付けたデータを蓄積することで，フィードバックの精度を向上させることが目的である．表情認識機能の詳細については [2] で述べている．

自然言語処理機能では，カウンセリング時に利用者との対話を行う対話機能と，文字ベースの悩み文のタグ付け機能が存在する．対話機能では，利用者との対話を行うことで，カウンセリングにおける情報収集や，先述の画像生成機能と合わせて利用者の緊張を解すことが目的である．そして，本論文で述べる FastText による文字ベースの悩み文のタグ付け機能では，カウンセリング全体を通して人工知能とのコミュニケーションの際利用者が入力した文を解析し，利用者がどのような悩みを抱えているのかのデータを蓄積することが目的である．

エージェント機能では，表情認識機能，自然言語処理機能によって得られた利用者の情報をまとめ，対話機能の会話内容や画像生成機能で生成されるキャラクターイメージの方向性等，システム全体の意思決定や優先度設定を行う．

3　悩みのジャンル分類システム

このシステムでは以下の流れを対話機能より利用者から入力された文章 (以下「悩み文」と表記) を受け取るごとに繰り返す.

1. 対話機能から悩み文を受け取る.
2. 悩み文を記録する.
3. 悩み文を解析し，どのような悩みを持っているかタグ付けを行う.
4. タグ付けした結果，今の会話でどのような悩みを利用者は抱えているかを示す評価データをエージェント機能に渡す.

エージェント機能に渡された評価データは，それ以降の会話内容や画像生成においてどのようなキャラクターイメージを生成するか等の意思決定に利用される．また，カウンセリング終了後に利用者に提示されるフィードバックにおいて，解析されたどの悩み文にどのような悩みの傾向を検出したかといった情報を含めることを考慮し，受け取った悩み文は記録している.

悩み文に付けられるタグは以下の 7 種類である．悩み文 1 つにつき 1 つのタグが付けられるのではなく，複数種のタグに渡る悩み文は該当する全てのタグが付けられる.

- ご近所の悩み
- 家族関係の悩み
- 学校の悩み
- 職場の悩み
- 生き方、人生相談
- 友人関係の悩み
- 恋愛相談

4　実験

4.1　データ収集

インターネットのお悩み相談サイトより，既に解決済みのお悩みを次のカテゴリごとに最新より 800 件ずつ取得を試みた．取得方法は Selenium を使用した Web スクレイピングであり，実際の取得件数は表 1 に記載する.

データの取得件数にばらつきがあるのは，スクレイピング時にデータを上手く取得できなかったものが存在するためである.

表 1: データの取得件数

	取得件数
ご近所	556
家族関係	568
学校	388
職場	500
人生相談	535
友人関係	413
恋愛相談	233
合計	3193

4.2　分類のための方法

収集したデータの質問文のみを抽出し，それを 8:2 の割合でランダムに学習データとテストデータへ振り分けた．各カテゴリのデータ件数は表 2 に記載する．振り分けたデータは MeCab を使い学習に適した形式に整え，FastText [3][4][5] にて学習を行った.

表 2: 各カテゴリのデータ件数

	学習データ	テストデータ
ご近所	435	121
家族関係	451	117
学校	314	74
職場	412	88
人生相談	439	96
友人関係	335	78
恋愛相談	183	50
合計	2569	624

4.3　分類の結果

学習を行ったモデルを用い，テストデータの予測を行った結果が表 3 である．カテゴリごとの正解率を表 4 に示す．また，実際に使用したテストデータより一部改変したデータと，その分類結果を図 1 に示す．これは，最初の__label__4 が正解のラベルで，次の__label__4 が分類されたラベル，その次の数字がこの悩み文が__label__4 である確率であり，最後に続く文章が MeCab によって分かち書きされた悩み文である.

4.4　考察

表 4 より，全体の正解率はおおよそ 6 割，カテゴリ別で見ると最も正解率が高いカテゴリがご近所の悩みカテゴリの 0.777 で，最も低いカテゴリが恋愛相談の 0.440 である.

表 3: テストデータ分類の結果

	ご近所	家族関係	学校	職場	人生相談	友人関係	恋愛相談
ご近所	94	6	1	8	9	3	0
家族関係	8	74	4	4	25	0	2
学校	0	3	48	8	6	6	3
職場	2	0	3	63	16	3	1
人生相談	13	6	4	12	51	4	6
友人関係	4	3	5	13	7	44	2
恋愛相談	0	5	2	4	10	7	22

__label__4 , __label__4, 0.9827991724014282, コンビニ で 初めて バイト を し て い ます 。 初回 の 出勤 が 夜 から で 、 スタッフ ルーム に 入り オーナー さん に こんばんは と 挨拶 を し たら 、 オーナー さん に おはよう ございます と 返さ れ まし た 。 特に 注意 等 は さ れ なかっ た の でし た が 、 バイト で は どの 時間 帯 でも おはよう ございます と 言う の でしょ う か ? 良けれ ば 教え て 下さい 。

図 1: 分類に使用したデータの例

表 4: カテゴリごとの正解率

	正解率
ご近所	0.777
家族関係	0.632
学校	0.649
職場	0.716
人生相談	0.531
友人関係	0.564
恋愛相談	0.440
合計	0.635

カテゴリごとに正解率に大きなばらつきがある要因として，以下のものを考える．

- 学習データ件数の差
- 悩み自体の類似性

学習データ件数の差について，表 1 の通り，最も正解率の低いカテゴリであった恋愛相談のデータ件数が他カテゴリと比較し 200〜300 件ほど下回っている．次いで件数が少なかった友人関係の悩みカテゴリの正解率も，同じく恋愛相談の次に低い 0.564 となっており，学習に用いたデータ件数とそのモデルの精度には相関関係があると考える．

悩み自体の類似性について，表 3 の家族関係の悩みカテゴリの分類結果を見ると，家族関係の悩みカテゴリのテストデータ件数が 117 件ある中，それを正しく分類できたのは 74 件である．ここで，正しく分類できなかったデータ 43 件の分類を見ると，他 6 カテゴリに均等に間違えたのではなく，人生相談カテゴリに間違えた件数が 25 件と明らかに集中している．これと同様のことは職場の悩みカテゴリ，人生相談カテゴリ，友人関係カテゴリ，恋愛相談カテゴリでも起きている．不正解のカテゴリに分類されると言っても，その分類される先のカテゴリは悩み文のカテゴリごとに傾向があるといえる．

これは，悩み文自体にそもそも正解カテゴリ以外のカテゴリを正解とする余地があったからだと考えられる．悩みは人それぞれであり，往々にして様々な問題が複雑に絡み合った結果であるため，それらは単一のカテゴリに分類しきれるものではないから，このような結果になったと予測する．

また，システム全体の今後の課題として，次のものがあげられる．

- 分類精度の向上
- 学習データ数の増強

分類精度の向上に関しては，現在データ件数の多いカテゴリは正解率も高くなるという結果が出ているため，このままデータ件数を増やしていくことで精度もある程度までは上昇していくと考えられる．

5 むすび

今回の実験は，最終的な目標である悩みのジャンル分類システムを構成する 1 機能についてのものでしかない．今後あげられる大きな課題の 1 つとして，複数種のタグに渡る悩み文は該当する全てのタグが付けられるとあるが，その複数種のタグに渡る悩み文をどのように判別するかというものがある．

今後はあげられた課題の解決の他，他システムへの組み込みを考えた際の細かな仕様の調整等を行うことを考える．

参考文献

[1] 西村 有紗，勝田 暁子，小湊 勇弥，清瀬 藍子，鈴木 和也，能戸 誓音，武藤 良，植木 一也: 人工知能でカウンセリングを身近に，HAI シンポジウム 2021 (2021)

[2] 小湊 勇弥，清瀬 藍子，西村 有紗，武藤 良，植木 一也: カウンセリングに必要な顔表情データの抽出，HAI シンポジウム 2021 (2021)

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov: Enriching Word Vectors with Subword Information, *Transactions of the Association of Computational Linguistics (TACL)*, Vol. 5, pp. 135–146. (2017)

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov: Bag of Tricks for Efficient Text Classification, *In Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2, pp. 427–431 (2017)

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, Tomas Mikolov: FastText.zip: Compressing text classification models, *In Proc. of International Conference on Learning Representations (ICLR)* (2017)