自由記述アンケート回答文の意味に基づく高速自動分析技術

東洋大学 総合情報学部 総合情報学科
教授　安達　由洋

令和2年9月15日

技術開発の背景

- 文科省はGIGAスクール構想の下で、１人１台のコンピュータの実現、高速大容量通信ネットワークの完備など、ハード・ソフト・指導体制の充実に向けての施策を推進している。
- タブレットやスマートフォンから入力された受講生の回答や意見をリアルタイムに表示するWebサービスが広く利用されている。

  サービス例：
  - アンケート回答文を入力された順に表示
  - アンケート回答文に対して、他の参加者が“星”を付ける
  - アンケート回答文に対する質問や意見などの応答をまとめて表示
  - 文中の単語やその部分列の照合により、アンケート回答文を分類

技術開発の動機

- 授業をしていると、学生が授業に集中していないことや、内容を理解できていないことに気づくことがある。
- より良い授業にするための改善点を知るにはどうすれば良いか。
- すぐに自由記述アンケートをとって、授業に集中できない理由、理解できない理由を聞けば良い。
- 自由記述アンケート回答文を、自然言語処理技術を用いて、超高速に分析する手法を開発したい。
- これができると、授業中に随時アンケートを取って、受講生の意見や理解状況を反映した授業を進めることができる。

技術のキーポイント

- 意味分散表現
  - 単語が表す意味情報を比較的低次元（数百次元）のベクトルで表現
  - テキスト中の単語について、その前後に出現する単語の情報を埋め込んでいる
- テキストをこの意味分散表現で特徴付けすると精度良く分析（分類・検索）できる
  - ベクトルが低次元で非常に高速に処理できる
  - 同義語、類義語、表記の揺れなどを適切に処理できる
  - 日本語Wikipediaから収取した網羅的な辞書が存在する

技術のキーポイント（続き）

- 単語の意味分散表現（ベクトル）を用いた計算

print(model.most_similar(positive=['王', '女'], negative=['男'], topn=1))

出力： [('妃', 0.766958177089691 2)]

print(model.most_similar(positive=['東京', '米国'], negative=['日本'], topn=1))

出力： [('ワシントンDC', 0.695015668869018 6)]

print(model.most_similar(positive=['名古屋市', '沖縄県'], negative=['愛知県'], topn=1))

出力： [('那覇市', 0.816728353500366 2)]

開発した技術

- 自由記述文の特徴を低次元ベクトルで意味分散表現して、話題や内容に基づき高速に分類・検索する技術を開発した。
  - 本技術を用いると、比較的短い10,000個の自由記述文を１秒以内に分析できる。
  - 同義語、類義語あるいは表記の揺れなどの問題が適切に処理される。
  - 本技術は汎用的で、多くの分野で有効に利用できる。

想定される用途

- 授業中に随時自由記述アンケートを取り、回答文を本技術を用いて即時に分析することで、受講生の意見や理解状況を反映した授業を進めることができる。
- 数千人が参加する講演会や集会などで、本技術を用いて即時に自由記述アンケート回答文を分析することで、参加者の意見や疑問などにリアルタイムに答える講演ができる。
- Webサービスで、本技術を用いて自由記述アンケート回答文やレビューを高速に自動分析することで、ユーザーの意見や意図を考慮に入れた高度で知的なサービスが提供できる。
- SNS上で発信される自由記述文を高速分析することで、最新の世論やトレンドなどを反映した商品開発などができる。

技術内容（1）

- 自由記述文を、意味分散表現して超高速に分析（分類・検索）する手法を開発

  処理１：各文を形態素解析して単語に分割する

  処理２：文を構成する各単語の分散表現ベクトルの相加平均をとって文ベクトルを計算する
  - 文の意味にあまり影響を与えない語からなる排除語集合を定義して、その集合に属する語の分散表現ベクトルを文ベクトルに含めない

  処理３：文ベクトルに基づいて、自由記述文を分類する
  - K-means++法（線形時間アルゴリズム）、Ward法、群平均法
  - クラスタラベル候補集合を定義して、コサイン類似度に基づき各クラスタのセントロイドに近いラベルをつける
  - 自由記述文の話題や内容に基づいた分類ができる

技術内容（2）

処理４：入力した問い合わせ文から、処理１～２で問い合わせ文ベクトルを求め、コサイン類似度により類似度の高い自由記述文を検索する

- 自由記述文またはキーワードで問い合わせができる
- 問い合わせ文の話題と内容に基づいた検索ができる

処理５：自由記述文を日本語評価極性辞書に基づいて、ポジティブ、ニュートラル、ネガティブのカテゴリに分類する

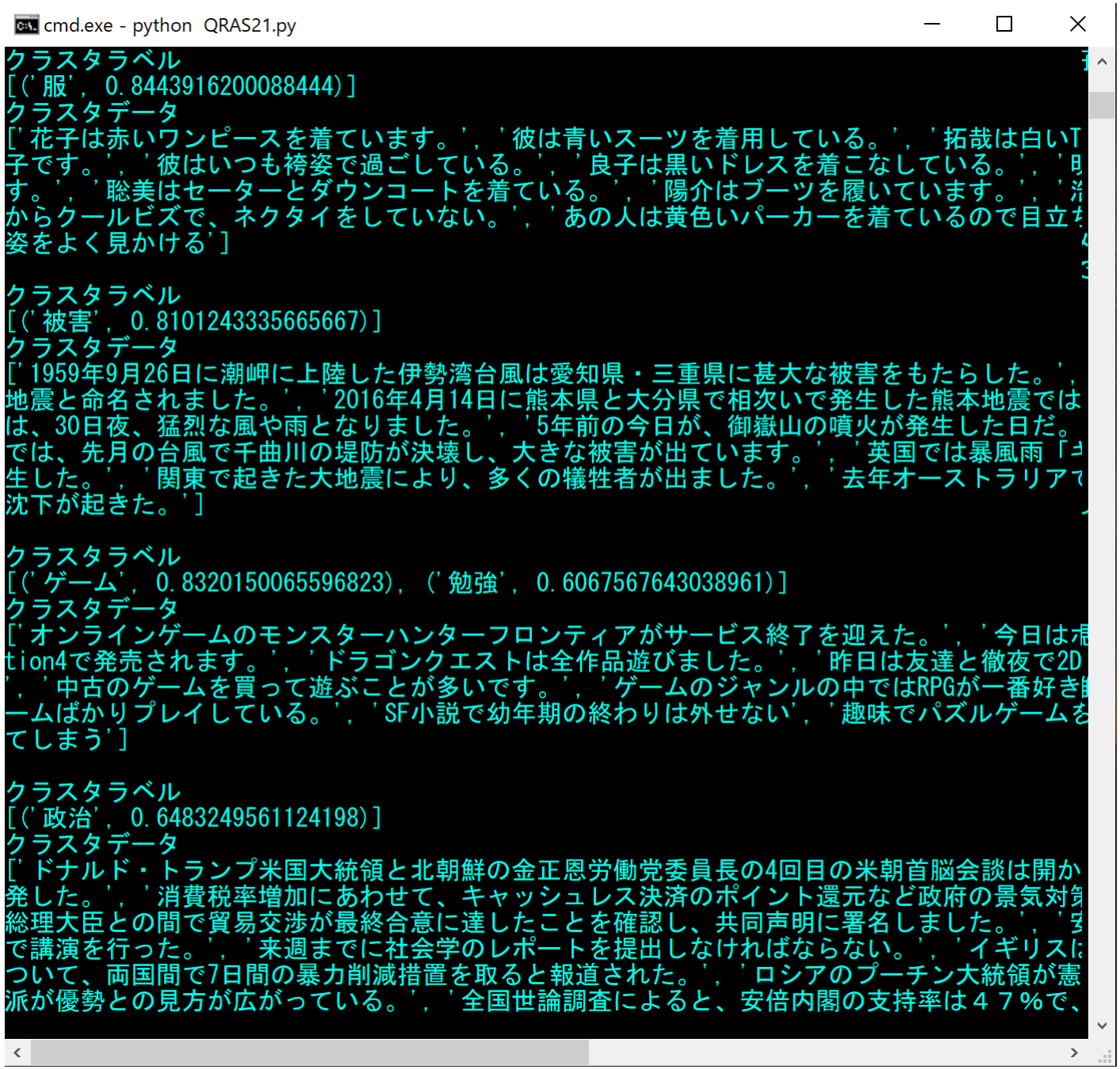
- 意味分散表現では下記2文は非常に近いベクトルとなり区別できない

  “入門プログラミングの講義が好きです。”

  “入門プログラミングの講義が嫌いです。”

- 極性評価により上記2文が分類できる

自由記述文分類の実行例



自由記述文分類の混同行列

- 150個の自由記述文データを用いて、kmeans++法により分類した結果の混同行列

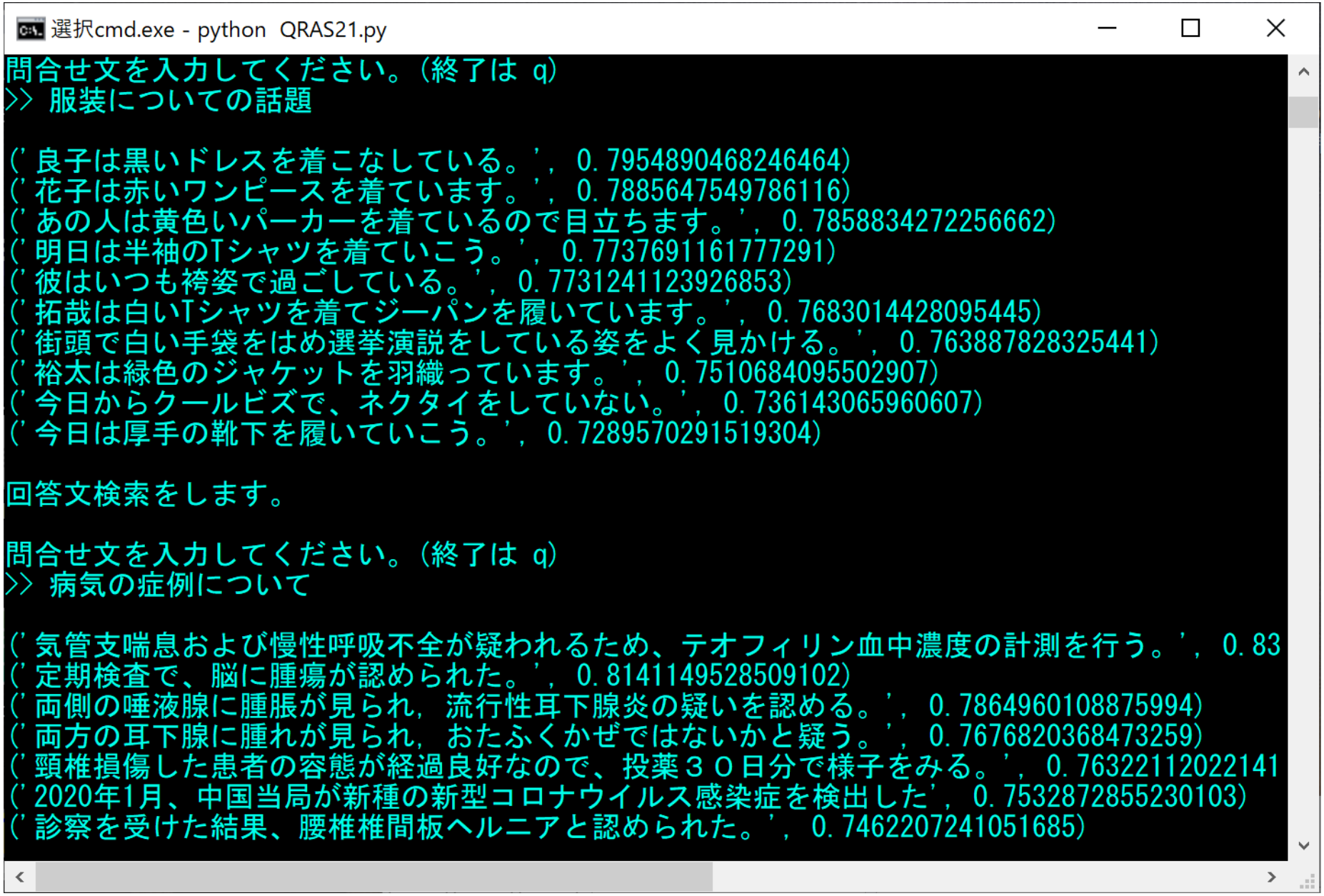
教師ラベル＼分類結果	勉強	動物	大学	天気	服	被害	ゲーム	政治	料理	小説	スポーツ	病気
勉強	7	1	1	2	0	0	0	1	0	0	0	0
動物	0	15	0	0	1	0	0	0	0	0	0	0
大学	0	0	10	0	0	0	0	0	0	0	0	0
天気	0	0	0	12	0	0	0	0	0	0	0	0
服	0	0	0	0	14	0	0	0	0	0	0	0
被害	0	0	0	0	0	11	0	1	0	0	0	0
ゲーム	0	0	0	0	0	0	12	0	0	0	0	0
政治	0	0	0	0	1	0	0	11	0	0	0	0
料理	0	0	0	0	0	0	0	0	15	0	0	1
小説	0	0	0	0	0	0	1	0	0	10	0	1
スポーツ	0	0	0	0	0	0	0	0	0	0	12	0
病気	0	0	0	0	0	0	0	0	0	0	0	10

自由記述文分類の精度

- 150個の自由記述文データを用いて、kmeans++法により分類した結果の精度

	適合度	再現率	F値
勉強	1.00	0.58	0.74
動物	0.94	0.94	0.94
大学	0.91	1.00	0.95
天気	0.86	1.00	0.92
服	0.88	1.00	0.93
被害	1.00	0.92	0.96
ゲーム	0.92	1.00	0.96
政治	0.85	0.92	0.88
料理	1.00	0.94	0.97
小説	1.00	0.83	0.91
スポーツ	1.00	1.00	1.00
病気	0.83	1.00	0.91

自由記述文検索の実行例



極性判定の混同行列

選択C:¥WINDOWS¥system32¥cmd.exe - python QRAS13.py

```
予測極性：　1
授業の進行が早くて話についていけませんでした
教師極性：　-1.0
予測極性：　0
演習を織り交ぜつつ授業が進行したので、よく理解できた
教師極性：　1.0
予測極性：　1
授業の進行がゆっくりすぎて退屈だった
教師極性：　-1.0
予測極性：　-1
先生の話が理路整然としていて、非常に分かりやすかった
教師極性：　1.0
予測極性：　0
授業内容と関係ない話が多くて、授業が間延びしている感じがする
教師極性：　-1.0
予測極性：　-1
心理学に興味があったので、とても面白そうだなと思いました
教師極性：　1.0
予測極性：　1
歴史は高校の時から苦手で嫌いです
教師極性：　-1.0
予測極性：　-1
黒板の文字が大きくて見やすかった
教師極性：　1.0
予測極性：　1
プログラミングの講義が好きです
教師極性：　1.0
予測極性：　1
プログラミングの講義が嫌いです
教師極性：　-1.0
予測極性：　-1
```

極性判定の混同行列

- 130個の極性判定用自由記述文データを用いた極性判定結果の混同行列

教師ラベル＼分類結果	ポジティブ	ニュートラル	ネガティブ
ポジティブ	46	12	1
ニュートラル	2	11	1
ネガティブ	4	17	36

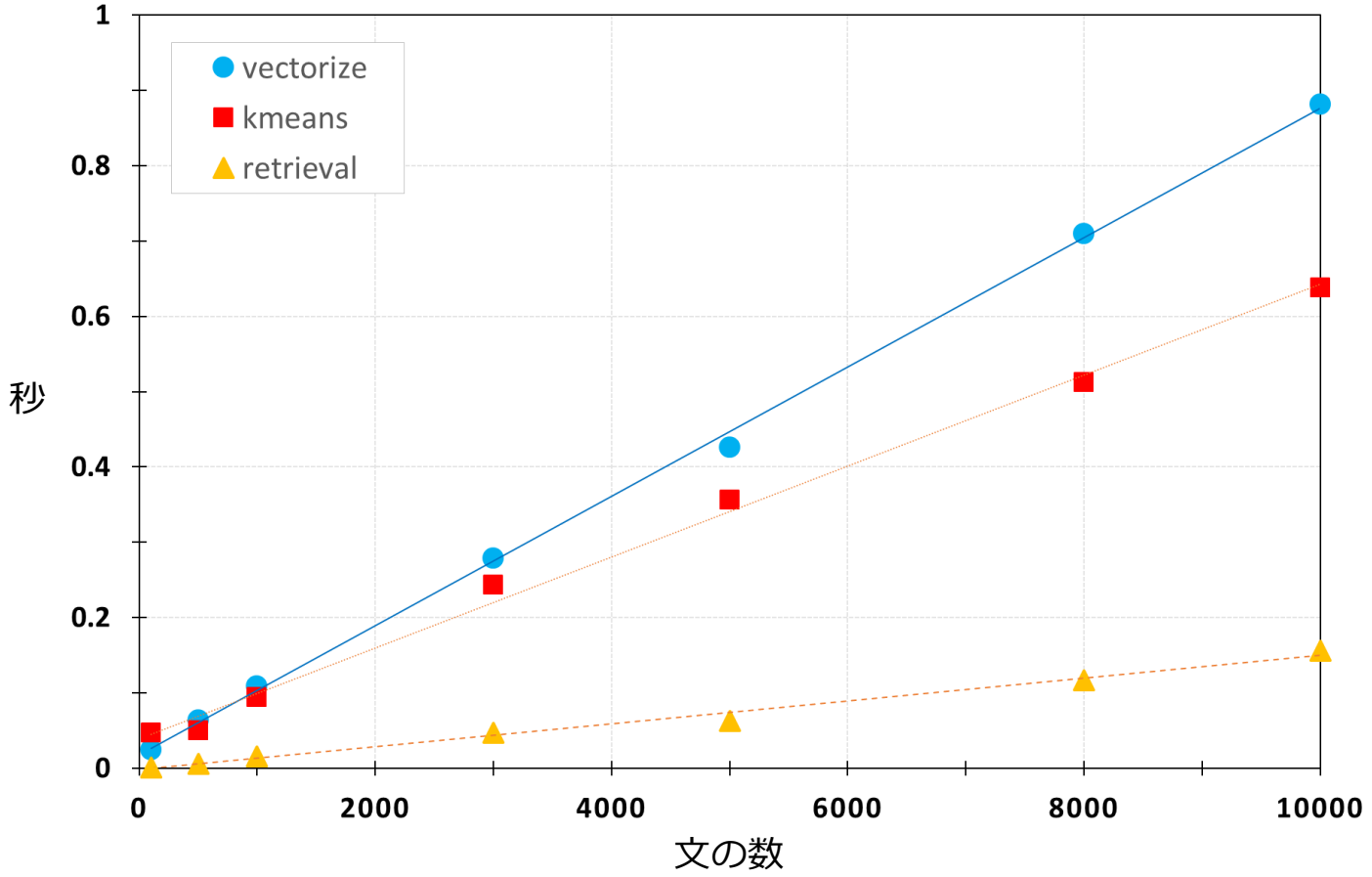
極性判定の精度

- 130個の極性判定用自由記述文データを用いた極性判定結果の精度

	適合率	再現率	F値
ポジティブ	0.88	0.78	0.83
ニュートラル	0.28	0.79	0.41
ネガティブ	0.95	0.63	0.76

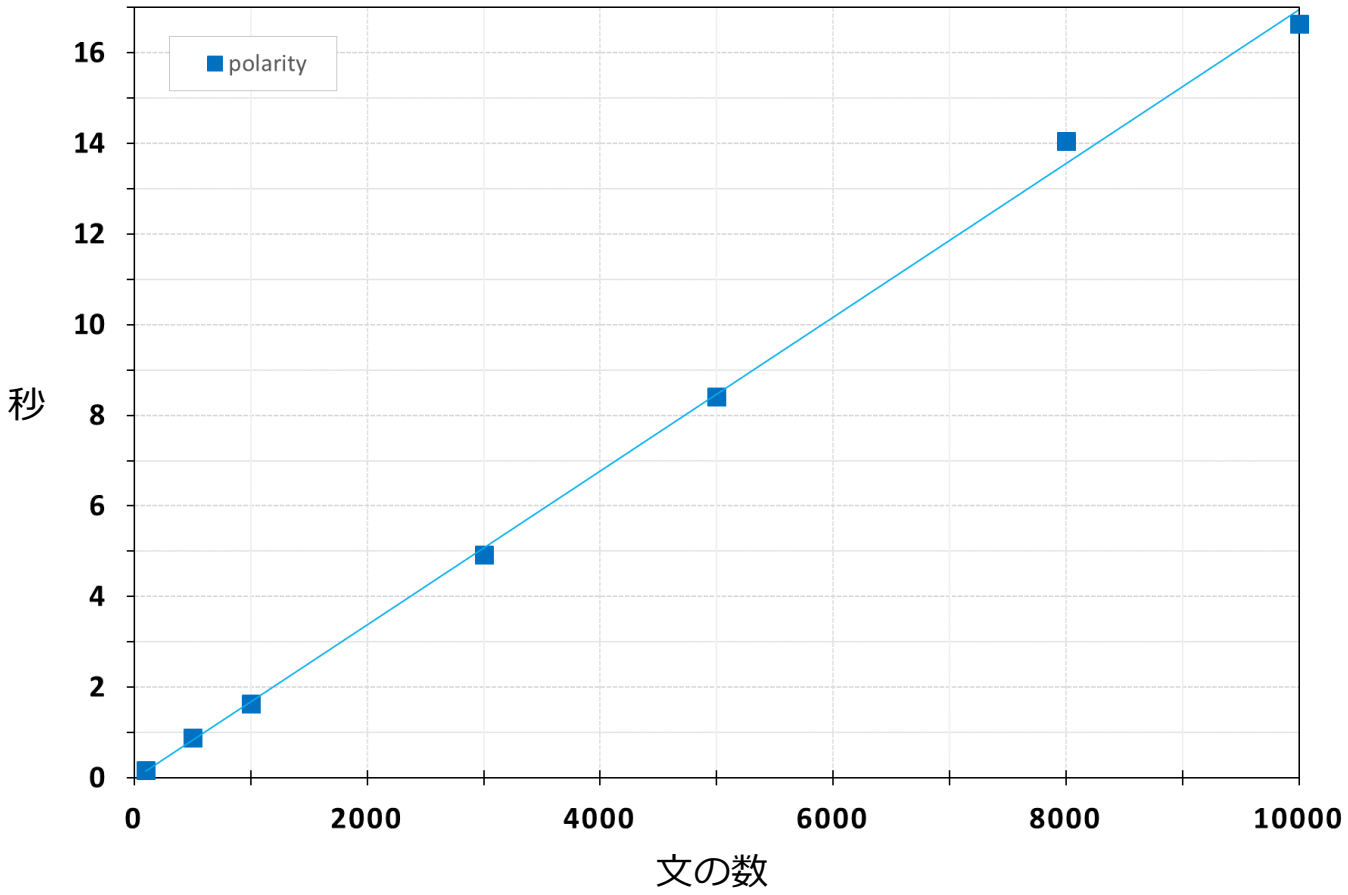
自由記述文分類の計算速度

- 10,000文の自由記述文に対して1秒以内に分類処理できる
  (CPU: Intel Core i7-6567U, Memory: 16GB)



極性判定の計算速度

- 3,000文の自由記述文に対して5秒以内に極性判定できる



企業への期待

本技術の導入が有効であると思われる企業：

- 小学校、中学校、高等学校、あるいは大学での授業環境を開発中の企業
- 大規模講演会や集会でのIT技術を駆使した対話環境を開発中の企業
- Webサービスの分野で高度で知的なユーザー対応機能を検討中の企業
  - 各商品の機能、特徴やアピール点を記述した説明文の中から、ユーザの問い合わせ文との類似度の高い商品を自動推薦する
- SNS上で発信される自由記述文を高速分析することで、世論やトレンドなどを反映した商品・サービスを開発する企業、あるいはそれをサポートする企業

本技術に関する知的財産権

- 発明の名称　：文を分類する方法、装置およびプログラム
- 出願番号　：特願2020-085337
- 出願人　：東洋大学
- 発明者　：安達由洋、根岸嵩典

お問い合わせ先

東洋大学
産官学連携推進センター
（研究推進部　産官学連携推進課）
ＴＥＬ　03-3945-7564
ＦＡＸ　03-3945-7906
e-mail　ml-chizai@toyo.jp