文書分類技術に基づくエントリーシートからの業界推薦

三王 慶太 [1]　酒井 哲也 [2]

新卒の就職活動においては，志望業界の選定が出来ずに苦労する就職活動生が多く見受けられる．就職活動生として主観的に新卒の就職活動を進めていると業界間の差は見え難いが，内定者として客観的に新卒の就職活動を振り返ると，少なからず志望者や内定者のパーソナリティに業界間の差が見て取れる．そこで，新卒の就職活動のエントリーシートにおいて，業界を問わず普遍的に採用されている設問を抽出し，その回答からパーソナリティを分析することで，志望者や内定者のパーソナリティの業界間の差を可視化する．さらに，明らかになった志望者や内定者のパーソナリティの業界間の差を利用し，就職活動生のエントリーシートをもとに業界推薦を行うシステムを設計する．今回設計した業界推薦システムは，ある程度似通った文書の分類を行い，同時にアノテーションを用いて情報を可視化することが求められた．ある程度似通った文書の分類には，文書の分散表現の次元が大きい方が都合が良く，アノテーションを用いた情報の可視化には，現実性の観点から，文書の分散表現の次元が小さい方が都合が良い．ここで生じたトレードオフの関係を解消する手法として，ADDV: Amplified Difference Document Vector を提案した．実際，推薦する業界の数を増やすことで，業界推薦の正確さ，及びアノテータによる定期的なアノテーション付けの容易さをある程度両立することが出来た．

1　はじめに

1.1　新卒の就職活動の現状

新卒の就職活動においては，志望業界の選定が出来ずに苦労する就職活動生が多く見受けられる．就職活動支援サイト「キャリタス就活」を運営する株式会社ディスコが実施した調査*1によると，2020 年卒業見込みの就職活動生 1,048 人のうち，2018 年 10 月 1 日時点で志望業界が明確に決まっていると回答したのは 19.8% に留まり，実に 80.2% もの就職活動生は志望業界を明確に定められていない．

また，同社が実施した調査*2によると，7 月 1 日の時点での 1 人あたりの平均エントリー社数は，2019 年卒業見込みの就職活動生で 30.7 社，2018 年卒業見込みの就職活動生では 39.6 社にも上った．エントリーする業界，企業の数が多くなると，各業界，各企業の理解が浅くなり，結果的に内定が遠ざかる．

今後 40 年働き続ける企業を探すための新卒の就職活動が，明確な意志を伴わずに行われ，結果的に内定をも遠ざけている現状は，危惧すべきであると言える．

1.2　志望業界を明確に定められない就職活動生のリスク

まず，一般的な新卒の就職活動の流れは，図 1 の通りである．

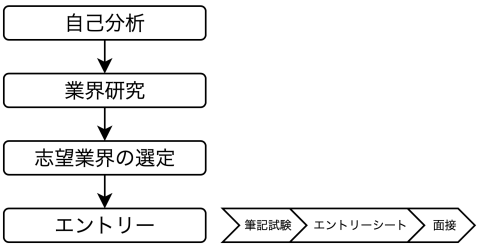


図 1　一般的な新卒の就職活動の流れ

中でも業界研究は，合同説明会，企業のセミナー，企業の冊子，IR 情報，OB 訪問などから得た情報を利用して行うため，情報を得るのに多大な時間を必要とする．志望業界の選定が出来ずに就職活動を始めた場合，多くの業界に対して業界研究が必要となるため，志望業界が明確に定まっている就職活動生と比較して，業界研究の質，量が低下する．

1.3　業界推薦システムについて

今回設計する業界推薦システムは，パーソナリティを軸として，客観的に適すると判断出来る業界に限定して業界推薦を行うことで，業界研究の対象を限定し，質，量共に高い業界研究を可能とすることに焦点を当てたシステムである．

志望業界の選定が出来ずに苦労している就職活動生は，業界推薦システムの出力に基づいて業界研究の対象を限定することで，業界研究の質，量を高めることが可能となる．一方，志望業界が明確に定まっている就職活動生に関しても，自身の強み，あるいは弱みを，その業界の志望者や内定者と比較しながら分析することが可能となる．

また，業界推薦システムがもたらす恩恵はそれだけに留まらない．業界推薦システムが出力した業界に対象を限定して業界研究を行うことで，業界，企業に対する理解の質，量を高めることが出来る．圧倒的な業界，企業理解は，面接官に対して入社意欲が高いことをアピールするための最良の手段であることから，業界推薦システムによって内定を近付けることが出来ると言える．

2　従来研究

2.1　心理学の知見を新卒の就職活動に応用した研究

心理学の知見を新卒の就職活動に応用した研究は，これまでにも多く行われてきた．古本は，自己 PR に如何なる要素が如何なる割合で盛り込まれるのかを分析した研究 [1] を行なった．小島は，大学生が就職活動初期に志望する働き方（職種，業種など）によって，エントリーシートへの記述に如何なる違いが生じるのかを分析した研究 [2] を行なった．飯田らは，学生と添削の専門家の間にある評価基準の差を分析し，その差を認識させることで，エントリーシートの記述レベルを上げる研究 [3] を行なった．

[1] 非会員　三井物産株式会社
keitasanno@ruri.waseda.jp
[2] 非会員　早稲田大学基幹理工学部情報理工学科
tetsuyasakai@acm.org

*1 https://www.disc.co.jp/wp/wp-content/uploads/2018/10/gakuseichosa_2020.pdf
*2 https://www.disc.co.jp/wp/wp-content/uploads/2018/07/19monitor_201807-1.pdf

2.2 情報処理技術を新卒の就職活動に応用した研究

情報処理技術を新卒の就職活動に応用した研究は，数が限られている．杉山らは，ユーザの行動履歴データに自然言語処理を施すことで，複数のエントリーの組み合わせを考慮した企業の分散表現を得る研究 [4] を行なった．坂元らは，企業のアピールポイントと学生の志望理由をもとにして，双方の関係性を分析する研究 [5] を行なった．

2.3 新規性

これらは，いずれも今回設計する業界推薦システムとは根本から焦点が異なる．企業のアピールポイントと学生の志望理由の対応を分析する坂元らの研究は，一見今回設計する業界推薦システムと似通っているようにも見えるが，企業のアピールポイントは，企業で働く社員が考えるものと，就職活動生が考えるものの間に乖離があるため，似て非なるものである．

3 提案手法

3.1 業界推薦の概要

今回設計する業界推薦システムは，業界を問わず普遍的に採用されている設問である，学生時代の経験，自己 PR を入力とする．入力に基づいて，パーソナリティを軸とした業界推薦，パーソナリティに関する特定の業界での強み，あるいは弱みの可視化を行う．ここで言うパーソナリティとは，就職活動生の性格，属性，志向などを指す．学生時代の経験，自己 PR などは，業界内で多用される特徴語（コンサル・シンクタンクにおけるクライアント，ソリューションなど）を含まず，志望者や内定者のパーソナリティを表す特徴語だけを含むので，ある程度似通っていると言える．そこで，志望者や内定者のパーソナリティの業界間の差を増幅して文書の分散表現を作成することで，業界推薦の正確さを上げることを目指す．

また，パーソナリティに関する特定の業界での強み，あるいは弱みを可視化するには，全ての特徴語に対して，アノテータが付けたアノテーションが必要となる．特徴語の数が膨大になると，年々変化する特徴語に対して，アノテータが定期的なアノテーション付けを行うのは非現実的になる．

ある程度似通った文書を分類する目的で，Offer Weight [6] [7]（以下，OW）を考慮して文書の分散表現に用いる小規模辞書を作成し，各分類に属する文書間の差を増幅した文書の分散表現を，**ADDV: Amplified Difference Document Vector**（以下，ADDV）と命名する[*3]．

今回設計する業界推薦システムは，ADDV を用いて業界推薦を行うことで，業界推薦の正確さ，及びアノテータによる定期的なアノテーション付けの容易さを両立することを目指す．

3.2 業界推薦の手順

業界推薦の手順の概要を図 2 に示す．

まず，取得したエントリーシートの中で，学生時代の経験，自己 PR などを含むエントリーシートだけを抽出する．なお，抽出したエントリーシートの集合を P とし，全ての $p \in P$ は，当該エントリーシートが提出された企業の業界分類へマッピングされているものとする．

図 2 業界推薦の手順の概要

全ての $p \in P$ に現れる単語の OW を，業界ごとに計算する．

OW に基づいて選定した特徴語を用いた，エントリーシートの分散表現を Entry Sheet Vector（以下，ESV）と呼ぶことにして，全ての $p \in P$ に対して ESV を計算する．

ESV の Centroid Vector を業界ごとに計算したものを Industry Vector（以下，IV）と呼ぶことにすると，当該業界の分散表現は IV に他ならない．

業界推薦の対象となるエントリーシートについても ESV を計算することで，IV との比較検討による業界推薦，及びパーソナリティに関する特定の業界での強み，あるいは弱みの可視化が可能となる．

なお，業界を推薦する手順の詳細は，3.3 節から 3.7 節で説明する．

3.3 パーソナリティが現れるエントリーシートの抽出

集合 U を取得したエントリーシート全体の集合とする．このとき，集合 U から学生時代の経験，自己 PR などを含む，パーソナリティが現れるエントリーシートの集合 $P(\subset U)$ を抽出する手順は，次の通りである．

ここで，ある単語 $t(i)$ を設問，あるいは回答に含むエントリーシートの集合を，次のように定義する．

$E(t(i))$ = $t(i)$ という単語を設問，
あるいは回答に含むエントリーシートの集合

すると，$E(t(i))$ の否定 $\overline{E(t(i))}$ は次のように定義される．

$$\overline{E(t(i))} = U \backslash E(t(i))$$

さらに，これらの表記法を用いて新たな集合を定義する．なお，$E($ プログラ $) \approx E($ プログラム $) \cup E($ プログラミング $)$ という近似が成立すると見做し，プログラム，あるいはプログラミングという単語を含むエントリーシートの集合を，プログラという単語を含むエントリーシートの集合として纏めた．

$S_1 = E($ 強み $)$

$S_2 = E($ 長所 $)$

$S_3 = E($ セールスポイント $)$

[*3] 以下，文書，エントリーシート，業界の分散表現と記載した場合には，ADDV による文書の分散表現のことを指し，単語の分散表現と記載した場合には，Word Embedding による単語の分散表現のことを指すものとする．

$S_4 = E(\text{学生}) \cap (E(\text{時代}) \cup E(\text{生活}))$

$S_5 = E(\text{大学}) \cap E(\text{時代})$

$S_6 = E(\text{人生})$

$S_7 = E(\text{PR})$

$S_8 = E(\text{アピール})$

$S_9 = E(\text{自信}) \cap \overline{E(\text{プログラ})}$

$S_{10} = E(\text{自慢})$

$$S_{11} = E(\text{経験}) \cap \overline{\left(\bigcup_{i=1}^{9} Q_i\right)}$$

$$\begin{cases} Q_1 = E(\text{クラブ}) \cup E(\text{サークル}) \\ Q_2 = E(\text{インターン}) \cup E(\text{アルバイト}) \\ Q_3 = E(\text{プログラ}) \cup E(\text{開発}) \\ Q_4 = E(\text{賞}) \cup E(\text{表彰}) \\ Q_5 = E(\text{IT}) \cup E(\text{ICT}) \\ Q_6 = E(\text{使用}) \cup E(\text{言語}) \\ Q_7 = E(\text{研究}) \cup E(\text{起業}) \\ Q_8 = E(\text{留学}) \cup E(\text{海外}) \\ Q_9 = E(\text{ボランティア}) \end{cases}$$

$S_{12} = E(\text{体験})$

$S_{13} = E(\text{エピソード})$

$S_{14} = (E(\text{力}) \cup E(\text{情熱})) \cap (E(\text{注}) \cup E(\text{入}))$

$S_{15} = E(\text{苦労})$

$S_{16} = E(\text{困難})$

$S_{17} = E(\text{挫折})$

$S_{18} = E(\text{壁})$

取得したエントリーシートの設問文に用いられている単語，設問文に対する就職活動生の回答を鑑みて，次のような集合 P を，就職活動生のパーソナリティが現れるエントリーシートの集合と定義する．

$$P = \bigcup_{i=1}^{18} S_i$$

集合 P を求めることは，就職活動生のパーソナリティが現れるエントリーシートを抽出することに他ならない．

3.4 業界ごとの OW の計算

全文書数を N，全文書中である単語 $t(i)$ が出現する文書数を n，ある条件に適合する文書数を R，R のうち，ある単語 $t(i)$ が出現する文書数を r とすると，ある単語 $t(i)$ が適合文書に共通する特徴語である度合いを計算する指標として，Offer Weight $OW(i)$ が定義出来る [8].

$$OW(i) = r \cdot \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}$$

各業界の志望者や内定者のパーソナリティが現れるエントリーシートの集合 P をデータセットとして採用し，全ての $p \in P$ に現れる単語の OW を業界ごとに計算する．これにより，全ての $p \in P$ に現れる単語が，各業界で志望者や内定者に共通するパーソナリティであると言える度合いを数値化することが出来る．

3.5 エントリーシートベクトル ESV の計算

全ての $p \in P$ で現れる単語のうち，IPA 品詞体系[*4]で品詞が名詞-一般，名詞-固有名詞-地域-一般，名詞-固有名詞-地域-国，名詞-サ変接続，名詞-形容動詞語幹のいずれかに分類される単語（以下，品詞要件を満たす単語）だけを抽出する．

最終的に ESV と IV の Cosine Distance を用いて業界推薦を行うため，文書の分散表現に用いる辞書に属する単語は，単語の分散表現が互いに乖離していることが望ましい．

そこで，品詞要件を満たす単語をクラスタリングし，X のクラスタに分類する．なお，評価実験のために行なったクラスタリングの詳細は，4.6 節で説明する．

ある単語 $t(i)$ の，ある業界 $b(k)$ での OW の順位（降順）が $r(i,k)$ であるとし，次が成立するものとする．

$$\{b(k) \mid k \in K\} = \text{全ての業界の集合}$$

品詞要件を満たす全ての単語に対し，次のように定義される OW の順位の最小値を求める．

$$r_{\min}(i) = \min\{r(i,k) \mid k \in K\}$$

OW の順位の最小値を求める例として，単語 $t(i)$，業界 $b(k)$，OW $OW(i)$（小数点以下切り捨て），OW の順位（降順）$r(i,k)$ が表 1 の通りである状況を想定する．

表 1　OW の順位の最小値の計算例

$t(i)$	$b(k)$	$OW(i)$	$r(i,k)$
留学	商社	523	3
留学	メーカー	179	251

この想定に基づいて OW の順位の最小値の計算をすると，$r_{\min}(\text{留学}) = 3$ となる．

さらに，各クラスタから OW の順位の最小値が当該クラスタ内で最小となる単語を求め，当該クラスタから特徴語として選定する．

互いに距離の離れた X のクラスタから 1 単語ずつ特徴語を選定するので，それらを文書の分散表現に用いる辞書とすることで，辞書内の単語の分散表現が互いに距離の離れた X 次元が定義出来て，各次元成分を TF-IDF で計算すると，全ての $p \in P$ を ESV に変換出来る．

3.6 業界ベクトル IV の計算

ESV の Centroid Vector を業界ごとに計算し，IV を求める．

例として，ESV の次元が $t(x)$，$t(y)$ からなり，各次元成分が表 2 の通りであるエントリーシートの集合 $P = \{p(1), p(2), p(3), p(4), p(5)\}$ を想定する．

この想定に基づいて IV の計算をすると，商社，及び IT・通信

[*4] http://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf

表 2　IV の計算例

文書	業界分類	[$t(x)$, $t(y)$]
$p(1)$	商社	[$TFIDF(x,1)$, $TFIDF(y,1)$]
$p(2)$	商社	[$TFIDF(x,2)$, $TFIDF(y,2)$]
$p(3)$	商社	[$TFIDF(x,3)$, $TFIDF(y,3)$]
$p(4)$	IT・通信	[$TFIDF(x,4)$, $TFIDF(y,4)$]
$p(5)$	IT・通信	[$TFIDF(x,5)$, $TFIDF(y,5)$]

の IV は，それぞれ次の通りである．

$$IV_{\text{商社}} = \left[\frac{\sum_{j=1}^{3} TFIDF(x,j)}{3}, \frac{\sum_{j=1}^{3} TFIDF(y,j)}{3} \right]$$

$$IV_{\text{IT・通信}} = \left[\frac{\sum_{j=4}^{5} TFIDF(x,j)}{2}, \frac{\sum_{j=4}^{5} TFIDF(y,j)}{2} \right]$$

3.7　業界推薦

例として，業界推薦の対象となるエントリーシート p' の ESV を，次のように想定する．

$$ESV_{p'} = [TFIDF(x,p'), TFIDF(y,p')]$$

ここで，分散表現 A，及び B の Cosine Distance を $\cos(A,B)$ と定義すると，$\cos(ESV_{p'}, IV_{\text{商社}}) < \cos(ESV_{p'}, IV_{\text{IT・通信}})$ であれば，エントリーシート p' に対しては IT・通信よりも商社を推薦すべきであり，逆に $\cos(ESV_{p'}, IV_{\text{商社}}) > \cos(ESV_{p'}, IV_{\text{IT・通信}})$ であれば，商社よりも IT・通信を推薦すべきである．このように，業界推薦の対象となるエントリーシートの ESV と，各業界の IV 間の Cosine Distance を測ることで，推薦すべき業界の順序を明らかにすることが出来る．

また，3.6 節で用いた $IV_{\text{商社}}$ の例と，$ESV_{p'}$ の同じ次元成分同士の差に関して，次の関係式が成立する状況を想定する．

$$\left| \frac{\sum_{j=1}^{3} TFIDF(x,j)}{3} - TFIDF(x,p') \right| << \left| \frac{\sum_{j=1}^{3} TFIDF(y,j)}{3} - TFIDF(y,p') \right|$$

このとき，エントリーシート p' で商社にエントリーする場合，$t(x)$ と比較して，$t(y)$ の特徴語に表されるパーソナリティが，商社の志望者や内定者に共通するパーソナリティとかけ離れていることが分かる．業界推薦システムが，このようにしてパーソナリティに関する特定の業界での強み，あるいは弱みを可視化することにより，就職活動生は，強みをさらにアピールする，弱みの解消方法を練るなどの，内定を近付ける合理的なアプローチを行うことが出来る．

最終的には，業界推薦の対象となるエントリーシートの ESV との Cosine Distance が小さい上位 Z 件の業界を推薦する．さらに，パーソナリティに関する特定の業界での強み，あるいは弱みを可視化することで，就職活動生の志望業界の選定を強力にサポートする．

4　実験と評価

4.1　データセット

過去に提出されたエントリーシートが企業別に閲覧出来る就職活動支援サービスのうち，エントリーシート掲載数が 20,000 を超えるのは，楽天株式会社が運営する「みんなの就職活動日記」[*5] と，株式会社ワンキャリアが運営する「ONE CAREER」[*6] だけである．クローラを設計して「みんなの就職活動日記」，及び「ONE CAREER」に掲載されたエントリーシートを取得し，業界推薦システムが用いるデータセットとして採用した．なお，業界推薦システムが用いるデータセットには，個人を特定出来るような情報は含まれない．

4.2　パーソナリティが現れるエントリーシートの抽出

3.3 節に則り，就職活動生のパーソナリティが現れるエントリーシートだけを抽出した．ただし，集合 M，集合 O を次のように定義し，全体集合 U を集合 M，集合 O の和集合とする．

M =「みんなの就職活動日記」
　から取得したエントリーシートの集合
O =「ONE CAREER」
　から取得したエントリーシートの集合
$U = M \cup O$

4.3　業界分類

次に，集合 P（就職活動生のパーソナリティが現れるエントリーシートの集合を P と 3.2 節で定義していた）を業界ごとに分類した．業界分類方法として，より詳細な業界分類が行われている「ONE CAREER」の業界分類方法を採用したため，「みんなの就職活動日記」から取得したエントリーシートに関しては，「ONE CAREER」の業界分類へのマッピングを行なった．「ONE CAREER」の業界分類方法は表 3 の通りである．

表 3　業界分類方法

業界分類	業界に属する企業の例
コンサル・シンクタンク	戦略コンサル，総合・IT コンサル
金融	銀行・証券，保険（生保・損保）
メーカー	消費財，自動車，医療機器・医薬品
商社	総合商社，専門商社
IT・通信	システム・ソリューション，情報通信
広告・マスコミ	広告，芸能・エンタメ（映像・音楽他）
人材・教育	人材（派遣・紹介），教育
インフラ・交通	電気・ガス・エネルギー
不動産・建設	総合不動産・デベロッパー，建設
旅行・観光	ホテル，旅行会社
ブライダル・美容・くらし	美容（フィットネス・エステ他）
医療・福祉	医療機関・調剤薬局，福祉
小売・流通	百貨店・スーパー・コンビニ
公務員・団体職員	中央省庁，独立行政法人，学校法人
その他	レストラン・フードサービス

[*5] https://www.nikki.ne.jp
[*6] https://www.onecareer.jp

取得したエントリーシート全体の業界分類別のエントリーシート数，及び集合 P に属するエントリーシートの業界分類別のエントリーシート数，パーソナリティが現れるエントリーシートの業界分類別の出現頻度（有効数字 2 桁で四捨五入した）は，表 4 の通りである．

表 4　業界分類別のエントリーシート数

業界分類	全体	$p \in P$	出現頻度
コンサル・シンクタンク	12,411	5,857	0.47
金融	67,269	22,266	0.33
メーカー	142,151	41,457	0.29
商社	15,838	6,347	0.40
IT・通信	50,411	13,735	0.27
広告・マスコミ	14,291	5,007	0.35
人材・教育	9,125	2,276	0.25
インフラ・交通	30,076	9,953	0.33
不動産・建設	21,494	6,940	0.32
旅行・観光	2,786	720	0.26
ブライダル・美容・くらし	1,210	264	0.22
医療・福祉	6,162	1,083	0.18
小売・流通	17,863	3,230	0.18
公務員・団体職員	1,240	714	0.58
その他	5,611	1,369	0.24

4.4　データセットの分割

業界推薦システムの正確さを評価するため，集合 P を 9 対 1 の割合で分割し，9 割をトレーニングデータセット，1 割をテストデータセットとして用いた．トレーニングデータセットの集合を P_{train}，テストデータセットの集合を P_{test} とする．

$$P = P_{\text{train}} \cup P_{\text{test}}$$

4.5　業界ごとの OW の計算

3.4 節に則り，全ての $p \in P_{\text{train}}$ に現れる単語の OW を，業界ごとに計算した．

4.6　エントリーシートベクトル ESV の計算

3.5 節に則り，$X = 300$（クラスタ数を X と 3.5 節で定義していた）として，K-Means 法（$K = 300$）を学習アルゴリズムとするクラスタリングを行い，全ての $p \in P$ の ESV を求めた．$X = 300$ としたのは，各クラスタに分類された単語が，一般的な日本語の文脈においてもある程度類似した単語であると判断出来る X の最小値（100 単位で X を増加させて判断した）が，$X = 300$ であったからである．なお，クラスタリングに用いる単語の分散表現は，Mikolov らの研究 [9] に基づいて開発された Word2Vec[*7] によって得た．

特徴語として選定した単語の分散表現が互いに乖離していることを確認するため，全ての $p \in P_{\text{train}}$ に現れる単語の Word2Vec による分散表現を，Principal Component Analysis [10]（以下，PCA）を用いて次元削減を施した後，特徴語の分散表現を True，非特徴語の分散表現を False として，2 次元空間にプロットした．プロット結果は，図 3 の通りである．

図 3　単語の分散表現

図 3 から，特徴語として選定した単語の分散表現は，互いに乖離していることが確認出来た．

4.7　業界ベクトル IV の計算

3.6 節に則り，全ての業界分類の IV を求めた．

4.8　業界推薦

3.7 節に則り，$Z = 1$（推薦する業界の数を Z と 3.7 節で定義していた）として業界を推薦した．

文書の分散表現に用いる辞書はクラスタリングによって求めたが，Word2Vec の学習アルゴリズムには，Continuous Bag-of-Words モデル（以下，CBOW）と，Skip-gram モデル（以下，SG）がある．Precision，Recall，F1 を Macro Average で計算すると，Word2Vec の学習アルゴリズム別の出力評価（有効数字 2 桁で四捨五入した）は，表 5 の通りである．

表 5　Word2Vec の学習アルゴリズム別の出力評価

学習アルゴリズム	Accuracy	Precision	Recall	F1
CBOW	**0.30**	**0.28**	**0.32**	**0.23**
SG	0.27	0.27	0.30	0.21

表 5 から，CBOW を学習モデルとしてクラスタリングを行う方が，出力評価が高いことが分かる．以下では，ADDV，あるいは ADDV を用いた手法と記述する場合，Word2Vec の学習アルゴリズムに CBOW を用いてクラスタリングを行った場合における提案手法のことを表すものとする．

ADDV を用いた出力の Confusion Matrix を，図 4 に示す．

図 4 から，ADDV は業界を問わず，ある程度の正確さで正しい業界推薦が出来ていることが分かる．

また，パーソナリティに関する特定の業界での強み，あるいは弱みを可視化するため，選定した特徴語にアノテーション付けを行なった．本論文の第一著者は，就職支援サービス「レクミー」[*8]

[*7] https://www.tensorflow.org/tutorials/representation/word2vec

[*8] https://www.recme.jp

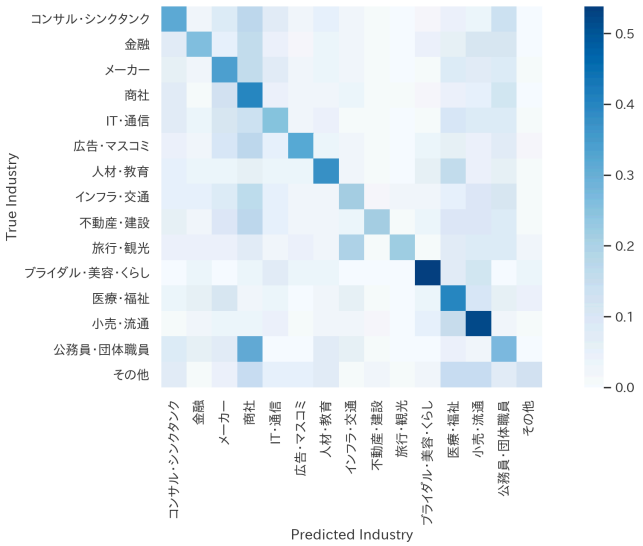


図 4　Confusion Matrix（ADDV）

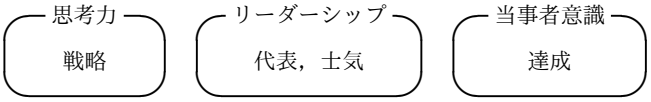
を運営する株式会社リーディングマーク[*9]にて，キャリアアドバイザーとして勤務している．そこで，特徴語に付けるアノテーションは，「レクミー」で採用されている評価指標に基づいたアノテーションである，思考力，リーダーシップ，当事者意識とした．

例として，ESV，及び IV の次元が戦略，代表，士気，達成で，各次元成分が表 6 である状況を想定する．

表 6　パーソナリティの分析例

ベクトル	[戦略 , 代表 , 士気 , 達成]
ESV	[0.35 , 0.70 , 0.20 , 0.65]
$IV_{\text{商社}}$	[0.50 , 0.40 , 0.60 , 0.45]
$ESV - IV_{\text{商社}}$	[-0.15 , 0.30 , -0.40 , 0.20]

また，特徴語のアノテーションは，次の通りであるとする．



この想定に基づいてパーソナリティの分析をすると，思考力，及びリーダーシップの分析値 v は，それぞれ次の通りである．

$$v_{\text{思考力}} = -0.15$$

$$v_{\text{リーダーシップ}} = \frac{0.30 + (-0.40)}{2} = -0.05$$

$$v_{\text{当事者意識}} = 0.20$$

求めた分析値 v が正であれば，そのアノテーションのパーソナリティは強みであり，逆に求めた和が負であれば，そのアノテーションのパーソナリティは弱みである．この想定の下では $v_{\text{思考力}} < 0$，$v_{\text{リーダーシップ}} < 0$，$v_{\text{当事者意識}} > 0$ が成立するので，当該エントリーシートで商社にエントリーする場合，思考力，リーダーシップは弱みとして改善すべきであり，当事者意識は強みとしてアピールすべきである．

このようにして分析したパーソナリティに関する特定の業界での強み，あるいは弱みを，業界推薦システムに出力させた．

4.9　機械学習を用いた手法との比較検討

ADDV を，機械学習を用いた手法と比較検討した．ただし，図 5 に示される通り，集合 P_{train} は不均衡データであるため，各業界に属するエントリーシート数に比例したコスト関数を用いて機械学習を行なった．

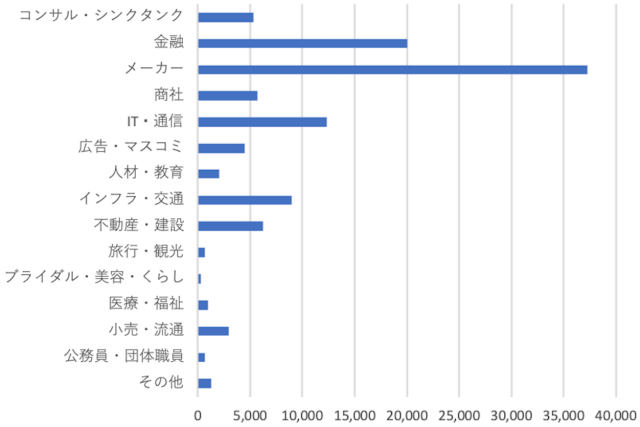


図 5　業界分類別のエントリーシート数（P_{train}）

機械学習を用いた手法では，品詞要件を満たす単語（3.5 節で定義していた）のみを対象に，半角・全角の統一，大文字・小文字の統一，数値の 0 への置換を施して辞書を作成した．

次に，Latent Semantic Indexing [11]（以下，LSI）を用いて次元削減を行い，エントリーシートの分散表現の次元を 300 次元まで下げた．

エントリーシートの分散表現と，当該エントリーシートが提出された企業の業界分類を教師有り学習させ，業界推薦の対象となるエントリーシートに対して，最も確信度が高い業界を推薦した．

機械学習を用いた手法として，Random Forest Classifier [12]（以下，RFC），及び Support Vector Machine [13]（以下，SVM）の出力評価を行なった．手法別の出力評価（有効数字 2 桁で四捨五入した）は，表 7 の通りである．ただし，Precision，Recall，F1 は Macro Average で計算するものとする．

表 7　手法別の出力評価

手法	Accuracy	Precision	Recall	F1
ADDV	0.30	0.28	0.32	0.23
RFC	0.45	**0.38**	0.19	0.22
SVM	**0.48**	0.35	**0.41**	**0.36**

RFC，SVM を用いた出力の Confusion Matrix を，それぞれ図 6，図 7 に示す．

図 4，図 6，図 7 から，ADDV と SVM の出力傾向は類似しているが，RFC は多くの入力に対してメーカーを推薦しており，

[*9] http://www.leadingmark.jp

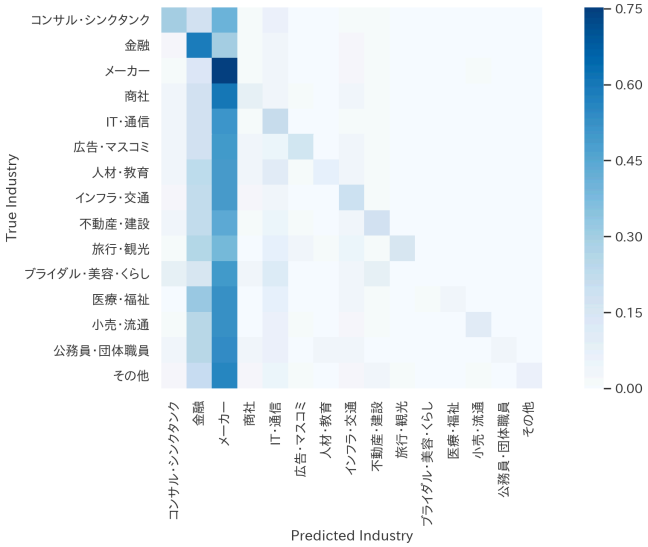


図 6　Confusion Matrix（RFC）

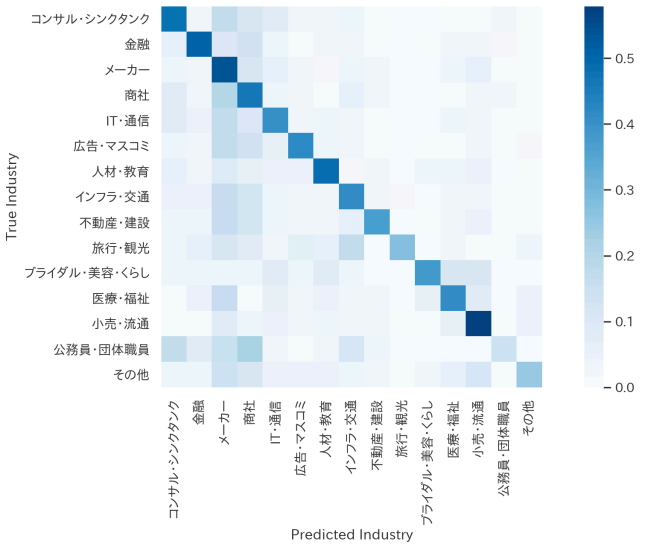


図 7　Confusion Matrix（SVM）

ADDV と SVM の出力傾向とは大きく異なることが分かる．

このことを，統計的仮説検定を用いて示す．Precision，Recall，F1 を，3 種類の手法，15 種類の業界に対応した，3×15 の行列で表す．3×15 の行列を用いて，Precision，Recall，F1 に関する Tukey HSD 検定 [14] [15] を行った結果（有効数字 2 桁で四捨五入した）は，表 8 の通りである．

表 8　Tukey HSD 検定による p 値

手法対	Precision	Recall	F1
RFC-ADDV	0.42	0.08	0.96
SVM-ADDV	0.61	0.24	0.07
SVM-RFC	0.94	**0.00**	**0.04**

表 8 から，SVM-RFC の手法対に関して，有意差が認められた．また，図 6 から，RFC を用いた出力は，その多くがメーカーを推薦しており，Accuracy は比較的高いものの，業界推薦システムとして適した出力であるとは言い難い．

これらの分析から，RFC は業界推薦システムとして適さないと判断し，以下では，ADDV，及び SVM の比較検討を行う．

エントリーシートの分散表現，及びその業界ごとの Centroid Vector である業界の分散表現は，LSI で次元削減を施すため，厳密には $c_i (1 \leq i \leq Y)$（LSI で次元削減を施す前に辞書内にある特徴語の個数が Y であったとする）を定数として，次のように単語の線型結合の形で表される．

$$\sum_{i=1}^{Y} c_i \cdot t(i)$$

機械学習を用いた手法においても，LSI で次元削減を施した後のエントリーシートの分散表現，及び業界の分散表現の各次元成分の差を計算し，当該次元を，LSI で次元削減を施す前に辞書内にあった単語の線型結合の形に戻すことで，ADDV と同様に，パーソナリティに関する特定の業界での強み，あるいは弱みを可視化出来る．

しかしながら，機械学習を用いた手法では，LSI で次元削減を施したとしても，辞書内にある全ての単語にアノテーションが必要である．手法別のアノテーションが必要な単語数は表 9 の通りである．

表 9　手法別のアノテーションが必要な単語数

手法	単語数
ADDV	**300**
RFC	21,189
SVM	21,189

オリンピックの開催，震災，コロナウイルスの蔓延などにより，各業界の特徴語は容易に変化し得る．表 9 に示される規模の年々変化する特徴語に対して，アノテータが定期的なアノテーション付けを行うことを考えると，現実性の観点では，アノテーションが 70 分の 1 以下で済む ADDV の方が，SVM よりも優れていると言える．

5　考察

5.1　併願業界数

第 1 志望業界に内定した 10 名に対し，併願業界数の調査を行なったところ，平均して 3.4 業界を併願していることが分かった．当該学生は，いずれも第 1 志望業界の大手一流企業[*10]に内定し，併願した他業界の大手一流企業[*11]からも多くの内定を得た，極めて優秀な学生（以下，模範的就職活動生）である．模範的就

[*10] 伊藤忠商事，キーエンス，トヨタ自動車，日本生命保険，パナソニック，プロクター・アンド・ギャンブル・ジャパン，丸紅，三井物産，三菱商事．
[*11] 全日本空輸，電通，デロイトトーマツコンサルティング，三井住友銀行，三井不動産，リクルートホールディングス（抜粋）．

職活動生は，最終的に本選考を経験した業界以外への理解も非常に深かった．

5.2 ADDV の出力評価の再考

改めて Z（推薦する業界の数を Z と 3.7 節で定義していた）の値を変えながら ADDV による業界推薦を行い，当該エントリーシートが提出された企業の業界分類を提示出来るかに関して，再度出力評価を行なった．ただし，5.3 節から，質，量共に高い業界研究が可能な業界数は，4 業界から 5 業界であると推測されるので，$1 \leq Z \leq 5$ の範囲での出力評価とした．推薦業界数別の ADDV の出力評価（有効数字 2 桁で四捨五入した）は，表 10 の通りである．

表 10　推薦業界数別の ADDV の出力評価

Z	Accuracy	Precision	Recall	F1
1	0.30	0.28	0.32	0.23
2	0.42	0.38	0.46	0.34
3	0.51	0.45	0.55	0.42
4	0.60	0.50	0.62	0.49
5	0.67	0.55	0.67	0.54

表 10 より，ADDV のネックとなっていた正確さは，Z の値を大きくすると，大幅に改善された．

5.3 業界推薦システムに求められる正確さ

今回設計した業界推薦システムは，業界研究の対象を限定し，質，量共に高い業界研究を可能にすることに焦点を当てたシステムである．模範的就職活動生は，平均して 3.4 業界を併願しており，最終的に本選考を経験した業界以外への理解も深いことから，質，量共に高い業界研究が可能な業界数は，4 業界から 5 業界であると推測される．

このことから，4 業界から 5 業界を推薦した出力にある程度の正確さが確認出来れば，業界研究の対象を限定し，質，量共に高い業界研究を可能にすることを目的とする業界推薦システムに求められる正確さの要件は満たされていると判断出来る．

6 結論と課題

6.1 結論

5.3 節に表 10 を照らし合わせると，ADDV の正確さは，SVM と比較すると見劣りするものの，業界推薦システムに求められる正確さの要件は満たすと判断出来る．また，4.9 節では，年々変化する特徴語に対してアノテータがアノテーション付けを行うことを考えると，現実性の観点から，アノテーションが 70 分の 1 以下で済む ADDV の方が，SVM よりも優れていた．

ここまでの結果から，ADDV を用いた業界推薦システムの有用性，及び SVM を用いた業界推薦システムと比較した場合の長所を示すことが出来たと言える．

6.2 課題

今回設計した業界推薦システムは，ある程度似通った文書の分類を行い，同時にアノテーションを用いて情報を可視化することが求められた．ある程度似通った文書の分類には，文書の分散表現の次元が大きい方が都合が良く，アノテーションを用いた情報の可視化には，現実性の観点から，文書の分散表現の次元が小さい方が都合が良い．ここで生じたトレードオフの関係を解消する手法として，ADDV を提案した．

実際，推薦する業界の数を増やすことで，業界推薦の正確さ，及びアノテータによる定期的なアノテーション付けの容易さをある程度両立することが出来た．しかし，この評価指標では，不正解の質（IV の距離が近い業界を推薦した結果の不正解であるのか，IV の距離が遠い業界を推薦した結果の不正解であるのか）を評価することが出来なかった．

現在，過去に提出されたエントリーシートが企業別に閲覧出来る就職活動支援サービスで，就職活動生の併願状況を確認できるサービスは存在しない．今後，進路として選んだ業界，併願した業界，業界の志望順位などの情報を付加したデータを蓄積することで，不正解の質も勘案する，より優れた業界推薦システムの完成に繋がるだろう．

参考文献

[1] 古本裕子: 就職活動における自己 PR 文の談話分析, 日本語教育方法研究会誌, Vol. 20, No. 1, pp. 80–81, 2013.
[2] 小島弥生: 就職活動におけるエントリーシートへの記述に関する探索的研究 - 志望する職種との関連の検討, 埼玉学園大学紀要 (人間学部篇), Vol. 10, No. 1, pp. 89–98, 2010.
[3] 飯田望美, 上野歩, 片山友昭, 中村亮太, 上林憲行: 就職活動におけるエントリーシートの作成を支援するサービスの検討 - 評価観点の一致と他者からの気づきを活用して, 第 74 回全国大会講演論文集, Vol. 2012, No. 1, pp. 585–586, 2012.
[4] 杉山裕貴, 雲居玄道, 後藤正幸, 桜井崇: 就職ポータルサイト上の行動履歴データに基づく企業の分散表現モデルに関する一考察, 研究報告数理モデル化と問題解決 (MPS), Vol. 2018-MPS-121, No. 11, pp. 1–2, 2018.
[5] 坂元哲平, 山下遥, 荻原大陸, 後藤正幸: 就職ポータルサイトにおける企業のアピールポイントと学生の志望理由のマッチング分析モデルに関する一考察, 情報処理学会論文誌, Vol. 58, No. 9, pp. 1535–1548, 2017.
[6] Robertson, S., Jones, K.: Relevance weighting of search terms, Journal of the American Society for Information Science, Vol. 27, No. 3, pp. 129–146, 1976.
[7] Robertson, S.: On term selection for query expansion, Journal of Documentation, Vol. 46, No. 4, pp. 359–364, 1990.
[8] Robertson, S., Jones, K.: Simple, proven approaches to text retrieval, University of Cambridge Computer Laboratory Technical Report, Vol. 356, pp. 3–8, 1994.
[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol. 2, pp. 3111–3119, 2013.
[10] Jolliffe, I.: Principal Component Analysis, International Encyclopedia of Statistical Science, Springer, pp. 1094–1096, 2011.
[11] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis, Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391–407, 1990.
[12] Breiman, L.: Random Forests, Machine Learning, Vol. 45, pp. 5–32, 2001.
[13] Cortes, C., Vapnik, V.: Support-Vector Networks, Machine Learning, Vol. 20, pp. 273–297, 1995.
[14] Tukey, J.: Comparing Individual Means in the Analysis of Variance, Biometrics, Vol. 5, No. 2, pp. 99–114, 1949.
[15] Sakai, T.: Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power, Springer, 2018.