概要

ウェブ上には数多くの様々な就職活動に関連する情報が存在する．これらの情報を自動で収集し有効活用することができれば便利である．そこで，本研究ではウェブ上の大量データから，就職活動に関連した情報を抽出し，分類する．

ウェブからの就職活動情報の自動抽出の関連研究に，沢の研究 [5] がある．沢の研究では，Yahoo!知恵袋のデータから就職関連記事を抽出し，分析している．本研究ではYahoo!知恵袋を含むより大きいウェブの情報源を用いるため，より多くの有用な情報が抽出できることが期待できる．また，抽出した就職関連情報を分類し，ユーザが役立てやすい形に近づける．

就職関連情報の抽出では，教師あり機械学習 (SVM)，ルールベース手法，ベースライン手法の 3 つの手法で性能を比較したところ，ルールベース手法が最良で F 値 7 割程度の性能を得た．就職関連情報の分類では，分類先として，“資格情報”，“職業情報”，“求職者ごとの情報”，“求人情報”，“関係無”，“就活支援情報”，“転職・再就職情報” の 7 つを設定した．教師あり機械学習 (SVM)，ルールベース手法，ベースライン手法の 3 つの手法で性能を比較したところ，ルールベース手法が最良で各分類先で F 値平均 6 割程度の性能を得た．

目 次

表 目 次

図 目 次

第1章　はじめに

就職活動は大切なことであり，就職活動を行う際に情報は不可欠である．そういった情報は様々な媒体からを得ることができ，ウェブ上にも多様で数多くの就職関連情報が存在している．これらは人手で収集するには多大な労力を必要とするため，自動で収集し有効活用することができれば便利である．そこで本研究では，ウェブ上の大量データから就職関連情報の抽出を行う．

ウェブからの就職活動情報の自動抽出の関連研究に，沢の研究 [5] がある．沢は Yahoo! 知恵袋のデータに，教師あり機械学習を用いカテゴリの付与をすることで，就職関連記事とその他の記事を分け，就職関連情報を抽出した．しかし，抽出された就職関連情報は分類などされていないものであり，ユーザが有効活用するには労力を必要とする．そこで本研究では，ユーザの負担を少なくするために，抽出された情報をより詳細に分類する．また，ウェブ上には Yahoo!知恵袋以外にも多くの情報源が存在しており，Yahoo 知恵袋より大きいウェブの情報源を利用することでより多くの有用な情報を抽出できることが期待できる．

本論文の主張点を以下にまとめる.

- 従来の就職関連情報の抽出の研究で用いられていた，Yahoo!知恵袋データよりも大量のデータを用い，就職関連情報の抽出を行った.
- 抽出された就職関連情報を分類し，ユーザが役立てやすい形に近づいた.
- 情報抽出において一般的な機械学習の他に，人手で作成したルールに基づくルールベース手法を用いた.

第2章　関連研究

情報抽出の手法，情報抽出の対象の 2 つの点に基づき，次節以降で関連研究を整理する．

2.1　機械学習やルールベースを用いた研究

堀の研究 [1] では，ウェブから変遷情報を抽出し，変遷情報の種類を分類した．ウェブから文データを抽出し，分類する点で，本研究とアプローチが似ている．堀は，ALAGIN の意味的関係抽出サービス [10] を用い変遷情報を抽出した．意味的関係抽出サービスを用い，ウェブからの情報の抽出を行い，抽出された結果を分類するというアプローチは，本研究と共通するアプローチである．堀の研究の流れを以下の図 2.1 に示す．

変遷情報の抽出では，抽出された変遷情報が真に変遷情報であるか判定をしている．ALAGIN の意味的関係抽出サービスでパターンを使用し抽出された文は，F 値が 0.86 と比較的高い性能で抽出されていた．更に，機械学習を組み合わせることで，F 値が 0.91 とより高い性能で変遷情報を抽出できることがわかった．変遷情報の種類の分類では機械学習を用い分類を行っている．機械学習を用いた分類では，いくつかの分類先では F 値が 6〜7 割程度の性能を得た．しかし，文の総数が少ないものでは性能が低い傾向にあった．また，一部の分類先が評価データ中に存在しなかったため，性能の評価ができなかった．文の総数を増やすことで改善が見込まれる．各分類先での性能を把握しておくことで，分類の性能の向上に役立てることができると考える．

端の研究 [3] では，ウェブ上からの文の抽出に機械学習を用い，抽出の性能が高かったことを示している．端は大量のウェブデータから感動を与える文を収集し，そういった文に多く含まれる単語を分析した．ウェブ上から感動を与える文を収集し，人手で感動を与える文か否かを判定したものを学習データとする．ウェブコーパスから取得した文を機械学習で自動推定し，得られた文が感動を与える文か否かを人手で判定し，学習データに追加する．これにより，機械学習で感動を与える文を抽出し，学習デー

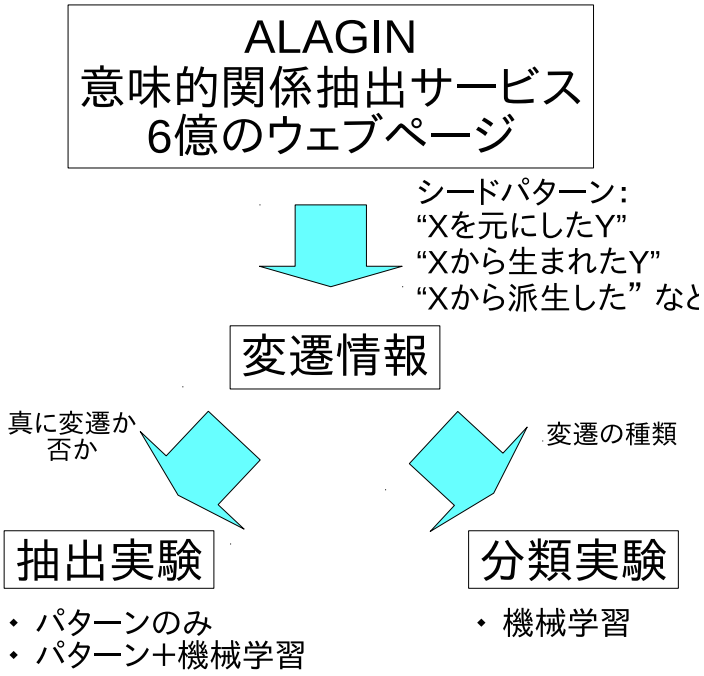

図 2.1: 変遷情報の抽出の流れ

タを追加した．続いて，学習データのうち，感動を与える文と判定されたものに多く含まれる単語を分析した．感動を与える文において，出現率 8 割以上かつ，出現頻度が 5 以上の単語を取り出している．結果，感動を与える文に多く含まれる単語として「人生」「人々」「幸福」「友情」「青春」「恋愛」などが得られた．最後に，自動抽出の性能を評価している．以下に述べる 4 つの手法でそれぞれ適合率，再現率，F 値を求めることにより性能を比較した．

1. 機械学習に基づく方法で，機械学習で抽出された学習データを n 回用いた場合の性能を求める．

2. 分析された単語を含む文を感動を与える文と判定する手法である．本研究のルールベース手法にあたる手法．

3. 「感動」という単語を含む文を感動を与える文と判定する手法．

4. 全ての文を感動を与える文と判定する，ベースライン手法．この手法で検出した正例の個数から，再現率の分母を推定している．

実験の結果，1 つ目の手法の機械学習で抽出された学習データを用いた手法の性能が一番高かった．機械学習で抽出された学習データを用いなかった時は適合率 0.06 だったが，機械学習で抽出された学習データを用いた時は適合率が 0.4 まで向上した．これにより，機械学習で抽出された学習データが有用だったことが示された．

ウェブからの情報抽出で機械学習だけでなくルールベース手法を取り入れ性能が向上した研究に，栗原ら [2] や高橋ら [6] の研究がある．

栗原らは，Twitter からの不具合情報の抽出で機械学習を用いたところ，適合率が 0.19 と非常に低かった．原因として，学習データの単語のガバレッジが不足していたことを挙げた．適合率を上げるには人手による正例の追加が検討されるが，高コストとなるため，現実的でないことを指摘している．そこでルールベース手法を利用し，人手で抽出のルールを作成したもので抽出を行った結果，適合率が 0.94 と大きく向上した．

高橋らは，職業の自動分類にルールベース手法と機械学習を合わせて利用することで，高い精度で分類できることを発見した．分類先は，仕事の内容，従業先事業の職種，従業上の地位，役職，従業先事業の規模を含む一連の解答群から被験者に自由回答で記述してもらい，総合的に判断し決定する．ルールベース手法では，職業コードに関する定義文や知識をルールとしてまとめた「職業辞書」を利用する．「職業辞書」に回答とマッチするルールがあればその職業コードを付与する．機械学習による手法では，素性に，「仕事の内容」に出現する単語，「従業先事業の種類」に出現する単語，「従業上の地位＋役職」を利用する．更に，機械学習とルールベース手法を組み合わせる手法では，以下に記述する 4 つを検討している．

1. ルールベース手法が出力した職業コードを素性に追加する
2. ルールベース手法でマッチしたルールを素性に追加する
3. ルールベース手法で出力した職業コードおよびマッチしたルールを素性に追加する
4. ルールベース手法が職業コードを決定できない場合に機械学習による方法の結果を利用する

実験の結果，4 つすべてにおいて，機械学習やルールベース手法を単独で利用するより高い正解率を得た．この内，ルールベース手法が決定した職業コードを素性とする方法が最も有効であった．

2.2 ウェブからの就職活動情報の自動抽出を行う関連研究

ウェブから就職関連情報の自動抽出を行った研究として，沢の研究 [5] や，前山らの研究 [4] が挙げられる．

沢は Yahoo!知恵袋データから就職活動情報を抽出し，分析した．Yahoo!知恵袋の記事のうち,「就活」「就職活動」を含む記事と，カテゴリが「就職活動」である記事を得た．得られた記事を分析し，就職関連情報の抽出に役立てた．記事分類に基づく分析では機械学習を利用し，元々カテゴリのない文書データにカテゴリを付与し，不要なカテゴリを除外することで就職関連情報を抽出した．実験の結果，恋愛のカテゴリは高精度で分類可能で，不要な記事の除外に役立った．特徴単語抽出に基づく分析では，上位 5 カテゴリごとに特徴単語を分析し，各カテゴリの内容を把握するのに役立てる．実験の結果上位 5 カテゴリの特徴的な名詞を多く抽出し，各カテゴリの内容の把握に役立った．

前山らはキーワード抽出を用いた就職活動支援システムを開発した．ユーザは指定ブラウザの拡張機能を設定することで，特定の就職支援サイトにアクセスした際，自動で対象の企業の情報 (企業名，企業の URL，就職支援サイトの URL) を取得できる．収集した企業の URL を基に就職支援サイトから企業情報を収集し，同業他社比較ページを作成する．また，収集した企業の URL を基に企業ホームページから，志望動機に適したキーワード抽出を行う．このキーワードに基づきツリー構造のキーワードグラフ (キーワードノード) を作成する．ユーザはキーワードグラフから，関連企業を参照することができる．このシステムにより，同業他社の比較と関連企業の発見が容易になり，志望企業決定までの作業の効率化が期待できる．

第3章　提案手法

3.1　提案手法で用いる基礎技術

3.1.1　ALAGIN 意味的関係抽出サービス

ALAGIN の意味的関係抽出サービス [10] は，パターンを入力することで，約 6 億のウェブページからパターンに適合した文と該当ページの URL を取得できるサービスである．以下の詳細な説明は,「意味的関係抽出サービスマニュアル」[11] を参考にしている．

このサービスでは「B に役立つ A」などのパターンを入力すると，これと合致した単語 A，B がウェブ文書から自動的に抽出される．いくつかの例を図 3.1，3.2 に示す．

シードパターン：「A が原因で起こる B」など

連鎖球菌 - 化膿性関節炎, EB ウイルス - 伝染性単核球症,
ツボカビ - カエルツボカビ症, 断層 - 直下型地震, 煤塵 - 環境問題,
フロン - 地球温暖化問題, トラウマ - PTSD,
ヒューマンエラー- 重大事故, 過冷却 - 結露, 窒素肥料 — 地下水汚染

図 3.1: 「原因-結果」関係インスタンスの例

ただし、特定の意味的関係に絞ったとしても，その知識は様々な言語パターンで書かれており，大量のインスタンスを獲得するには大量の言語パターンが必要という問題がある．それらを人手で用意する作業は高コストとなる．

このサービスでは人手コストを最小限にするため，少数の言語パターン (シードパターン) を入力するだけで稼働するように設計されている．これを可能にしているのが，シードパターンと同じ意味的関係を表す，一種の言い換えとなる言語パターン (類似パターン) を自動学習する機能である．類似パターンの学習は，同じインスタンスを獲得できるパターン同士とは良い言い換えであるという考えに基づいている．例をあ

シードパターン：「A を防止する B」など

情報漏えい - 暗号化ソフトウェア, 不正アクセス - ファイヤーウォール機能,
床ずれ - エアマット, 鳥害 - 防鳥ネット, 手荒れ - ラノリン,
老化 - ガラクタン, 壁内結露 - 羊毛断熱材, 尿モレ - 立体ギャザー,
2 白とび - ND フィルター, 腐食 - クロームメッキ

図 3.2: 「トラブル-予防策」関係インスタンスの例

げると,「A が B の原因になる」「B の原因である A」を入力すると，これらと同じインスタンスを獲得しやすい「A によって起こる B」「A で B が発生」「B を招く A」など，多くの人がすぐには思いつきにくい言語パターンも含め，大量の類似パターンを学習する．最終的には学習された前類似パターンを用いて大量のインスタンスを獲得する．

サービスの基本的なデータ量を以下の図 3.3 に示す.

抽出対象の文書数：約 6 億ウェブページ
対象とする単語数：約 100 万
利用可能な言語パターン：約 58,700,000 種類

図 3.3: ALAGIN 意味的関係抽出サービス 基本的なデータ量

3.1.2 機械学習 (*SVM*)

SVM は教師あり機械学習の一つで，空間を超平面で分割することにより，2 つの分類からなるデータを分類する手法である [7]．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン) が大きい (図 3.4 参照) ほどオープンデータで誤った分類をする可能性が低くなると考えられ，このマージンを最大にする超平面を求め，それを利用し分類する．一般的には，上記の他の方法に「ソフトマージン」と呼ばれる学習データでマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，線形分離ができない問題に対応するために，表平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，式 3.1 識別関数を用い分類することと等価であり，その識別関数の出力値の正負によって 2 つの分類を判別できる [8],[9].

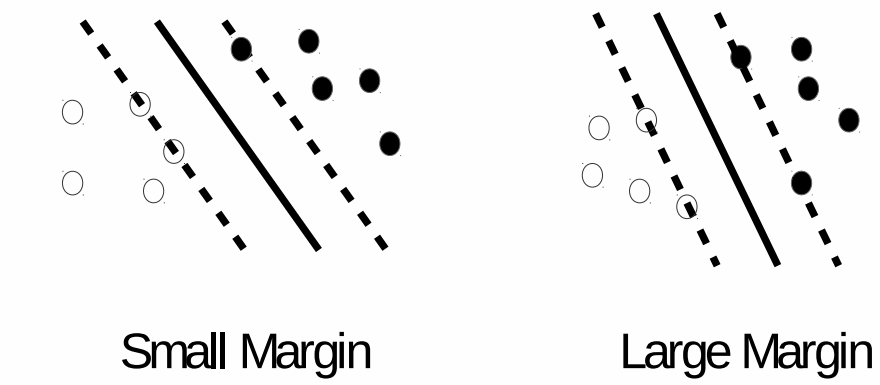

図 3.4: マージン最大化

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \\
b &= -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2} \\
b_i &= \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}
\tag{3.1}
$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i (i = 1, ..., l, y_i \in \{1, -1\})$ は学習データの文脈と分類先を意味し，関数 sgn は，

$$
\begin{aligned}
sgn(x) = \quad & 1 \quad (x \geq 0) \\
& -1 \quad (otherwise)
\end{aligned}
\tag{3.2}
$$

であり，また，各 α_i は式 (3.4) と式 (3.5) の制約のもと式 (3.3) の $L(\alpha)$ を最大にする場合のものである．

$$
L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{3.3}
$$

$$
0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{3.4}
$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.5}$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本論文では式 (3.6) の多項式のものを用いる．

$$K(\mathbf{x}, \mathbf{y}) \quad = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.6}$$

C, d は実験的に設定される定数である．本論文ではすべての実験を通して C,d はそれぞれ 1 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (3.1) の和をとっている部分はこの事例のみを用いて計算される．つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない．

サポートベクトルマシン法は 2 値分類器であるため，分類が 3 個以上のデータを扱う際ペアワイズ手法を組み合わせ利用している [8]．ペアワイズ手法とは，N 個の分類を持つデータの場合，異なる二つの分類先のあらゆるペア (N(N-1)/2 個) を作り，各ペアごとにどちらが良いかを 2 値分類で求め，最終的に N(N-1)/2 個 の 2 値分類の分類先の多数決により，分類先を求める方法である．本研究では 2 値分類しか用いないため，ペアワイズ手法等を用いない．

3.1.3 10 分割クロスバリデーション

就職関連情報の抽出の実験では，10 分割クロスバリデーションを用いる．

10 分割クロスバリデーションでは，事例を 10 個に分割し，そのうちの 1 つをテストデータとし，残りを学習データとする．10 分割された事例それぞれをテストデータとし，10 回推定を行う．例を以下の図 3.5 に示す．10 回分の結果を組み合わせ，テストデータ全体の推定結果を得る．

3.1.4 ルールベース手法

本研究におけるルールベース手法では，単語辞書に含まれる単語と同一の単語が入力された文中に含まれるとき，正例と判定する．単語辞書に含まれる単語をルールベー

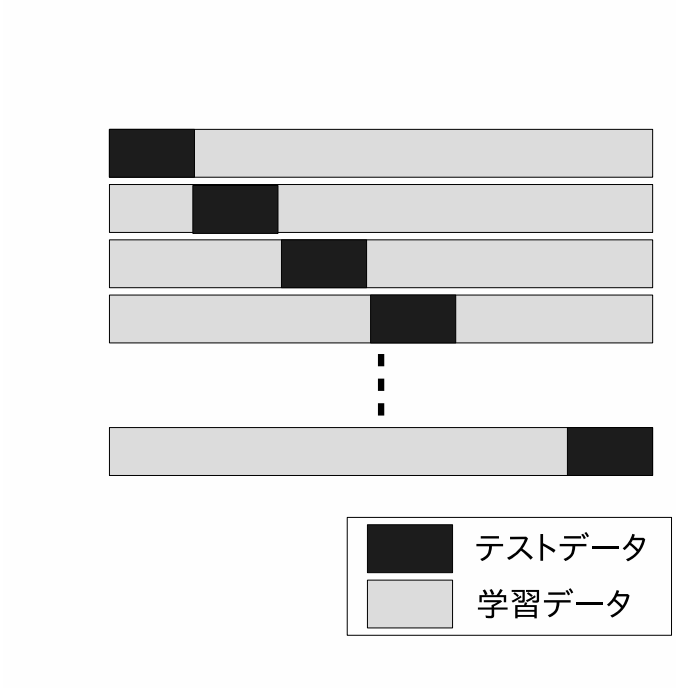

図 3.5: 10 分割クロスバリデーション

ス指定単語という．本研究では手法の評価を公平に行う目的で，評価データとは別のデータでルールベース指定単語を作成する．ルールベース手法での評価の例を図 3.6 に示す.

3.1.5 適合率，再現率，F 値

手法の性能を示す，適合率 (P)，再現率 (R)，F 値 (F) の求め方を以下に示す.

T：本来の正例の個数
T ’：その手法での正例の個数
$T \wedge T$ ’：本来の正解と，その手法での正解の重なりの個数

$$P = \frac{T \wedge T'}{T'} \tag{3.7}$$

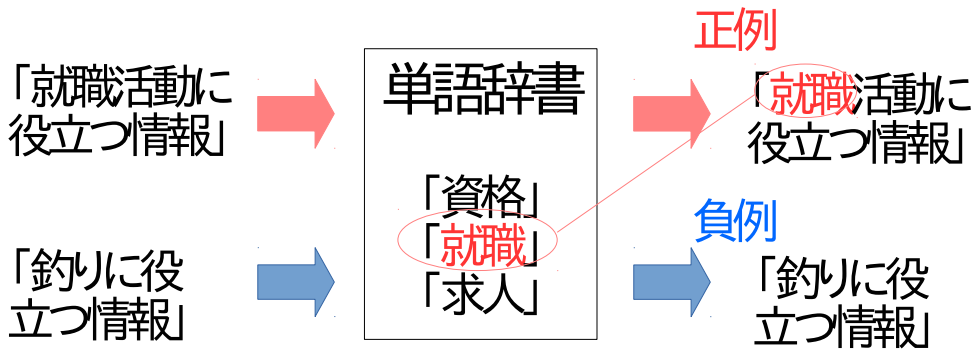

図 3.6: ルールベース手法

$$R = \frac{T \wedge T\,'}{T} \tag{3.8}$$

$$F = (\frac{\frac{1}{P} + \frac{1}{R}}{2})^{-1} \tag{3.9}$$

F 値は，適合率と再現率を加味し，その手法の性能を表す値である．

3.2 就職関連情報の抽出

ウェブ上の大量データから情報を抽出するために，ALAGIN の意味的関係抽出サービスを利用する．ALAGIN の意味的関係抽出サービスでは，パターンを入力し，文と該当ページの URL が得られる．そのため，就職関連情報である文だけでなく，URL 先に有益な就職関連情報があると期待できるような文も就職関連情報とする．

就職活動に役立つ情報を取得したいため，“B に役立つ A” という意味関係をシードパターンとした．シードパターンと，類似パターンの一部を図 3.7 に示す．

これらの意味関係を含む文集合を取得する．得られた文集合から就職活動に関係するものを，教師あり機械学習やルールベース手法により取り出す．全ての文章を就職関連情報と判定する，ベースライン手法も取り入れた．

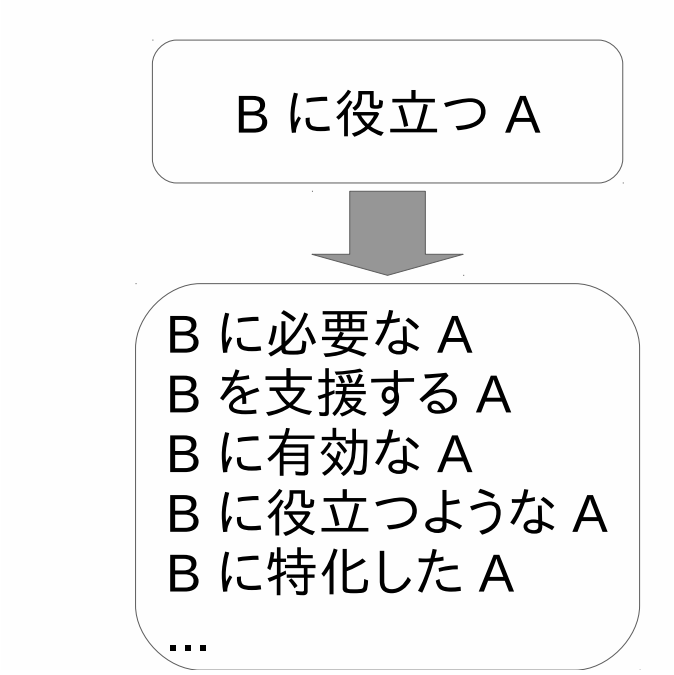

図 3.7: シードパターンと類似パターン

機械学習では，正解の分類先を付与したデータを作成し，学習データとして用いる．機械学習で用いる素性は，文を ChaSen[12] を用い単語に分割し，その単語を利用する．機械学習には SVM を用い，10 分割のクロスバリデーションで評価する．

ルールベース手法では,「資格」「求人」「就職」を含む文を就職関連情報として抽出する．これらのルールベース指定単語は事前に評価データとは別のデータにおいて人手で定める．

ベースライン手法は全てを正例と判定する手法である．ベースライン手法で検出した就職関連情報の個数から再現率の分母を推定する．

3.3 就職関連情報の分類

就職関連情報の抽出で最も性能が高かった手法で文を取り出し，分類する．分類は学習データを使用し，類似した文集合を人手でまとめることで，以下の 7 つに決定した．

資格情報

資格に関する情報を含む文

例：“小型船舶操縦士の資格取得には、二通りの方法があります。”

職業情報

特定の職業に関する情報が含まれるとされる文

例：“医療事務の就職で大切なのは、年齢だけではなく、レセプト作成などの専門知識と、病院の顔としての人柄や心配りです。”

求職者ごとの情報

誰に向けられた就職関連情報なのかが明確な文 (学生，主婦，障害者など)

例：“本書は、視覚障害者の大学進学から就職までの円滑な学生生活の実践的ノウハウを、視覚障害学生のニーズに沿って詳細に解説した、視覚障害学生との相互理解が深まるガイドブックです。”

求人情報

求人に関する情報が含まれる文

例：“地元の情報誌などを購入し、求人情報を調べる。”

関係無

就職関連情報でないと判定される文

例：“僕ら二人は大学で所属していたゼミから就職先まで同じ。”

就活支援情報

就職活動に必要な知識や技術に関する情報が含まれる文

例：“履歴書類の書き方、就職のための法律知識、ビジネスマナー、面接の受け方など。”

転職・再就職情報

転職や再就職に関する情報が含まれる文

例：“女性転職に有利な資格として考えられるのは、ホームヘルパーや医療事務、調剤報酬請求事務、介護事務等の福祉系のもの、又はパソコン関係の資格や経理関係の資格等も女性転職に有利な資格だと言え　…”

各文が上記分類先に分類されるか否かを，教師あり機械学習やルールベース手法より判定し，性能を比較する．ひとつの文に複数の分類先を付与する場合がある．

機械学習で用いる素性文を ChaSen を用い単語に分割し，その単語を利用する．機械学習には SVM を用いる．本研究では,「各分類先」と「その他の分類先」で 7 回に分け 2 値分類を行う．このような 2 値分類の例を図 3.8 に示す.

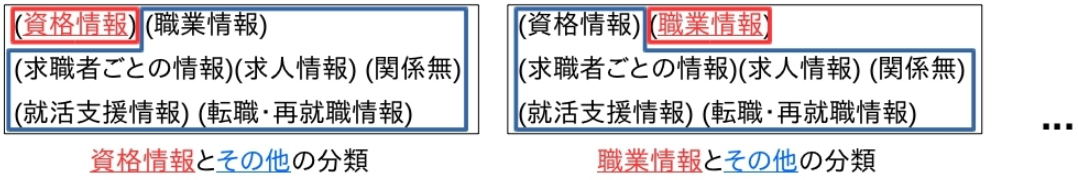

図 3.8: 2 値分類

ルールベース手法では，各分類先ごとに単語辞書を作成し，ルールベース指定単語が含まれる文をその分類先とする．ルールベース指定単語の一覧を表 3.1 に示す．ルールベース指定単語は学習データを人手で分析し作成する．

表 3.1: ルールベース指定単語一覧

分類先	ルールベース指定単語
資格情報	技術士資格，日商簿記，建築士，建設業経理事務士，カラーコーディネーター，販売士　他 676 件
職業情報	ネットワーク設計，作業療法士，生産管理，フラワーデザイン，IT 観光コーディネータ，マーケティング，証券，営業アシスタント　他 1,253 件
求職者ごとの情報	学生，大学生，新卒，既卒，中途，高校生，生徒，在学，進路，進学，留学，主婦，外国人，障害者，保護者，成績，院生，留学生，大学院生，大卒，大学新卒，中途採用，第二新卒，大学卒，既卒者，院卒，大学院卒，短大卒，新規学卒，修士卒，新卒大学生，高校卒，第二新卒者，高校新卒者，第 2 新卒，卒業予定者，高卒者，新規学卒者，新卒学生，専門学校卒，大学新卒者，新規卒業者，専門卒，高卒予定者，大卒者，新規高卒者，学卒者，新卒予定者，高校卒業予定者，新規高等学校卒業者，卒業見込，卒業見込み，修了見込み，退職予定，修了予定，高校三年，大学院，短大，子育てママ，子育て主婦，障がい者，父兄，内申点，新卒者，高卒，大学卒業者，高校卒業生，高校卒業者，新規学校卒業者，卒業予定，高等学校卒業者，大学卒業生，大学卒業予定者，高等学校卒業予定者，就職希望者，外国人留学生，短大生，高専卒，中卒，女子学生
求人情報	求人，求職，応募資格，活かし，活かす，仕事探し
就活支援情報	自己表現，志望動機，職業訓練，マナー，セミナー，インターンシップ，ワープロ，面接，スキルアップ
転職・再就職情報	転職，再就職

ベースライン手法は，全てを各分類先と判定する手法である．ベースライン手法で各分類先に分類された文章の個数から再現率の分母を推定する．

第4章　実験

4.1　就職関連情報の抽出

就職関連情報を抽出する際，“B に役立つ A” などの意味関係を含む文集合を ALAGIN の意味的関係抽出サービスで取得し，746,432 文が得られた．各手法の性能を調べるためにそのうちからランダムに抜き出した 1,000 文を評価に利用した．就職関連情報の抽出の実験の流れを図 4.1 に示す．

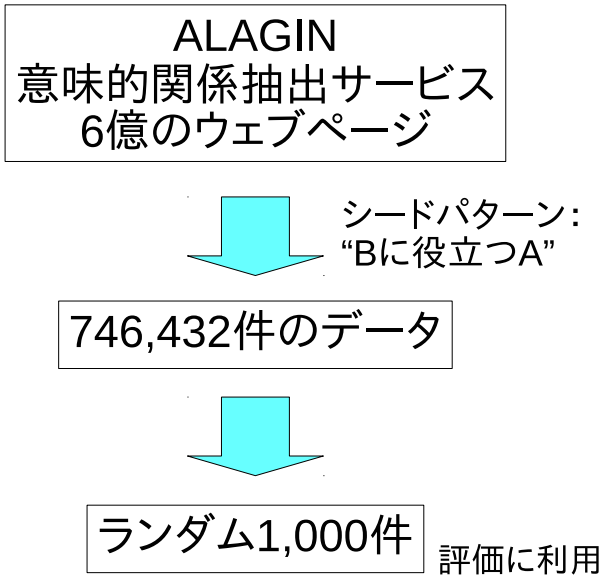

図 4.1: 就職関連情報の抽出実験の流れ

教師あり機械学習の実験データはこの 1,000 文を用いて 10 分割クロスバリデーションで評価した．ルールベース手法では,「資格」「就職」「求人」を含むものを就職関連情報とした．全ての文を就職活動関係と判別するベースライン手法も利用した．

各手法での結果を表 4.1 に示す．

表 4.1: 就職関連情報の抽出結果

手法	適合率	再現率	F 値
教師あり機械学習	0.75	0.39	0.51
ルールベース手法	0.70	0.74	0.72
ベースライン手法	0.03	1.00	0.06

ルールベース手法が他手法よりも性能が良く，F 値 7 割程度の性能を得た．

4.2 就職関連情報の分類

就職関連情報の抽出実験ではルールベース手法が最も性能が高かった．ゆえに，ルールベース手法で就職関連情報を抽出し，分類した．就職関連情報の抽出実験でルールベース手法を利用すると，9,908 文が得られた．そのうちランダムで抜き出した 300 文を，就職関連情報の分類の評価データに利用し，別の 300 件を学習データとした．就職関連情報の分類実験の流れを以下の図 4.2 に示す．

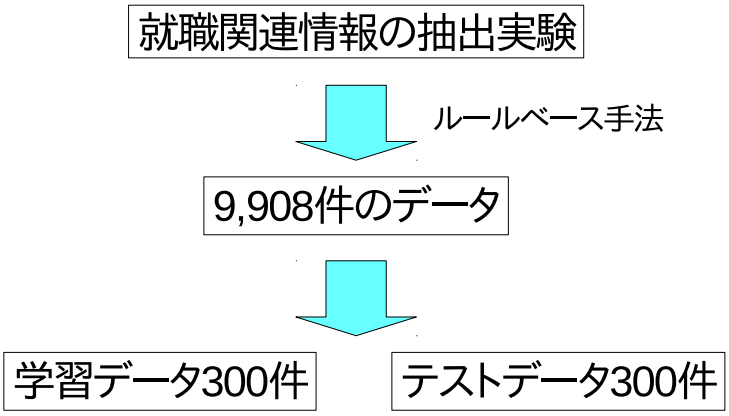

図 4.2: 就職関連情報の分類実験の流れ

評価データ 300 文中の，各分類先の出現数を表 4.2 に示す．

就職関連情報の抽出の実験同様，教師あり機械学習，ルールベース手法，ベースライン手法での性能を比較する．手法の結果を表 4.3，4.4，4.5，4.6 に示す．

ルールベース手法で F 値平均 6 割となっており，機械学習の F 値平均 5 割より高かった．また，ベースライン手法では，“資格情報”，“求人情報”，“転職・再就職情報” の

表 4.2: 就職関連情報の分類先

分類先	出現数	分類先	出現数
資格情報	94	関係無	91
職業情報	73	就活支援情報	56
求職者ごとの情報	22	転職・再就職情報	35
求人情報	39		

表 4.3: 就職関連情報の分類先と F 値

分類先	機械学習	ルールベース	ベースライン
資格情報	0.75	0.83	0.48
職業情報	0.52	0.50	0.39
求職者ごとの情報	0.34	0.53	0.14
求人情報	0.69	0.75	0.23
関係無	0.45	0.47	0.47
就活支援情報	0.34	0.44	0.31
転職・再就職情報	0.45	0.89	0.21
平均	0.51	0.63	0.32

分類先で F 値約 8 割程度の性能が得られており，ある程度うまく分類することができることがわかった.

表 4.4: 機械学習での各分類先の適合率，再現率，F 値

分類先	適合率	再現率	F 値
資格情報	0.82 (65/79)	0.69 (65/94)	0.75
職業情報	0.60 (33/55)	0.45 (33/73)	0.52
求職者ごとの情報	0.37 (7/19)	0.32 (7/22)	0.34
求人情報	0.72 (26/36)	0.67 (26/39)	0.69
関係無	0.50 (37/74)	0.41 (37/91)	0.45
就活支援情報	0.62 (13/21)	0.23 (13/56)	0.34
転職・再就職情報	0.67 (12/18)	0.34 (12/35)	0.45
平均	0.61	0.44	0.51

表 4.5: ルールベース手法での各分類先の適合率，再現率，F 値

分類先	適合率	再現率	F 値
資格情報	0.73 (91/125)	0.97 (91/94)	0.83
職業情報	0.48 (39/81)	0.53 (39/73)	0.51
求職者ごとの情報	0.42 (16/38)	0.73 (16/22)	0.53
求人情報	0.63 (36/57)	0.92 (36/39)	0.75
関係無	0.30 (91/300)	1.00 (91/91)	0.47
就活支援情報	0.69 (18/26)	0.32 (18/56)	0.44
転職・再就職情報	0.8 (35/44)	1.00 (35/35)	0.89
平均	0.58	0.78	0.63

表 4.6: ベースライン手法での各分類先の適合率，再現率，F 値

分類先	適合率	再現率	F 値
資格情報	0.31 (94/300)	1.00 (94/94)	0.48
職業情報	0.24 (73/300)	1.00 (73/73)	0.39
求職者ごとの情報	0.07 (22/300)	1.00 (22/22)	0.14
求人情報	0.13 (39/300)	1.00 (39/39)	0.23
関係無	0.30 (91/300)	1.00 (91/91)	0.47
就活支援情報	0.19 (56/300)	1.00 (56/56)	0.31
転職・再就職情報	0.12 (35/300)	1.00 (35/35)	0.21
平均	0.20	1.00	0.32

第5章　今後の課題

本研究では，ウェブ上の大量データから抽出した就職関連情報を分類し，いくつかの分類先で 8 割程度の性能で分類することができた．しかし，これを前山ら [4] の研究のように実用化するには，より精度を上げる必要がある．また，これらのデータをユーザに提示するとすれば，その提示の方法も考える必要がある．

また，7 つの分類先を利用したが，これらの分類先は更に細分化することが可能と考える．例えば，職業情報は更に職種ごとにカテゴライズすることが考えられる．

自動分類の性能を上げるという観点では，高橋ら [6] が行ったように，ルールベース手法と教師あり機械学習を合わせて利用し，性能を向上させる方法が考えられる．

上記に上げた点を考慮し，今後の課題として，ルールベース手法と機械学習手法を合わせて利用することで自動分類の性能を上げる方法を検討するほか，分類の詳細化や，ユーザへの提示手法の改善が求められる．

第6章　おわりに

本研究では就職関連情報の抽出と分類を行い，その両方でルールベース手法が教師あり機械学習より性能が高かった．

就職関連情報の抽出では，教師あり機械学習がＦ値5割に対し，ルールベース手法では7割の性能が得られた．この実験結果は,「資格」「就職」「求人」を含む文が就職関連情報であることが多かったことを示している．

就職関連情報の分類では，教師あり機械学習がＦ値平均5割に対して，ルールベース手法では平均6割の性能が得られた．就職関連情報の分類実験ではルールベース手法にて，“資格情報”，“求人情報”，“転職・再就職情報” の分類先で8割程度の性能が得られた．

謝辞

本研究を進めるにあたり，終始に渡り研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科計算機工学 C 講座の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授，徳久雅人講師に心から御礼申し上げます．その他様々な場面で御助言を頂きました計算機工学 C 講座研究室の皆様方に感謝の意を表します．

参考文献

[1] 堀 さな子. パターンと機械学習を用いた大規模テキストからの変遷情報の抽出と分類. 言語処理学会第 19 回年次大会, pp.592-595, 2013.

[2] 栗原 光平, 嶋田 和孝 ルールと機械学習を用いた Twitter からの不具合情報の抽出. 電子情報通信学会, 言語理解とコミュニケーション研究会 (NLC), NLC2014-1, pp.1-6, 2014.

[3] 端 大輝, 村田 真樹, 徳久 雅人. 感動を与える文の自動取得と分析. 言語処理学会第 18 回年次大会, pp.303-306, 2012.

[4] 前山 侑平, 安留 誠吾. キーワード抽出を用いた就職活動支援システム. 情報処理学会創立 50 周年記念 (第 72 回) 全国大会, pp.“ 4-853 ”-“ 4-854 ”, 2010.

[5] 沢 真之介. ウェブからの就職活動に関する情報の抽出. 鳥取大学工学部知能情報工学科卒業論文, 2014.

[6] 高橋 和子, 高村 大也, 奥村 学. 機械学習とルールベースの組み合わせによる職業コーディング. 情報処理学会研究報告. 自然言語処理研究会報告, pp.53-60, 2004.

[7] 村田 真樹, 井佐原 均. 機械学習を用いた日本語格解析 — 教師信号借用型と非借用型. 情報処理学会自然言語処理研究会 2001-NL-144, pp.113-120, 2001.

[8] Taku Kudoh. “TinySVM: Support Vector Machines”, http://www.chasen.org/taku/software/TinySVM/, 2000.

[9] Nello Cristianini, John Shawe-Taylor. “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods”, Cambridge University Press, 2000.

[10] 高度言語情報融合フォーラム ALAGIN 意味的関係抽出サービス: https://alaginrc.nict.go.jp/.

[11] 高度言語情報融合フォーラム: “意味的関係抽出サービスマニュアル”, https://alaginrc.nict.go.jp/, (独) 情報通信研究機構 MASTAR プロジェクト言語基盤グループ.

[12] ChaSen: http://chasen-legacy.sourceforge.jp/.