

修士論文

就職相談システムのための
求人票中の文字列のデータベース化

平成24年度修了
三重大学大学院工学研究科
博士前期課程 電気電子工学専攻

重永 宜也

目次

第1章	はじめに	1
1.1	研究の背景	1
1.2	提案システムの概要と本論文の位置づけ	2
1.3	現在の問題点と従来研究の概要	4
1.4	本研究の目的及び概要	6
第2章	研究対象とする文書画像	8
2.1	求人票	8
2.2	使用する求人票	8
2.3	求人票の文字認識結果	14
第3章	求人票画像から企業データベースへの自動登録	16
3.1	企業データベースの概要	16
3.2	統一フォーマットへの変換	17
3.2.1	文字列整理	18
3.2.2	内容候補の整列と抽出	19
3.3	項目名候補と対応する内容の抽出	24
3.4	内容の選択	27
第4章	評価実験	29
4.1	実験概要	29
4.2	提案法による内容の抽出精度についての評価	29
4.2.1	評価用文書画像	29
4.2.2	評価方法	29
4.2.3	結果と考察	30
4.3	手入力の場合と比較した提案法の運用効率についての評価	35
4.3.1	評価用文書画像と入力対象項目	35
4.3.2	評価方法	35

4.3.3	結果と考察	36
4.4	本手法の問題点	39
第5章	おわりに	41
5.1	本論文のまとめ	41
5.2	今後の課題	41
	謝辞	43
	参考文献	44
	発表論文リスト	46

図一覧

1.1	就職相談システムの使用イメージ	3
1.2	就職相談システムの開発過程	4
1.3	授業科目と職種の関係性の表（一部）	4
1.4	多様性の吸収モジュール	7
2.1	全体が罫線による表構造	10
2.2	一部が罫線による表構造	11
2.3	文字のみによる求人票	12
2.4	白地以外の背景のある求人票（画像はグレースケール）	13
2.5	求人票（入力画像）と文字認識結果の例	15
3.1	企業データベースのテーブル構造	16
3.2	企業データベースへの自動登録手順	17
3.3	統一フォーマットの例（2行のCSV形式）	18
3.4	文字認識結果（上）に対する文字列整理処理の適用結果（下）	19
3.5	内容判定用キーワード辞書	21
3.6	住所の階層構造モデル	22
3.7	別の行に記述された2つの内容が結合した例	23
3.8	内容候補間の分離処理の適用結果	23
3.9	対応する項目名が表記されていない例	24
3.10	内容候補に対する項目名の付加処理の適用結果	24
3.11	項目名判定用キーワード辞書	25
3.12	求人票における項目名表記	25
3.13	内容の表記に注目した内容候補の判定手順	26
3.14	項目名候補と内容候補を抽出した文字列	28
3.15	項目名候補が競合した例	28
3.16	統一フォーマット（CSV形式）	28
4.1	内容の抽出の成功例1	30
4.2	内容の抽出の成功例2	31

4.3	内容の抽出の成功例 3	31
4.4	内容の抽出の失敗例（文字認識処理による影響）	33
4.5	内容の抽出の失敗例（項目名の誤判定）	34
4.6	内容の抽出の失敗例（内容の抽出数の不足）	35
4.7	求人票画像から企業データベースへの自動登録システムの将来像	39
4.8	複雑な表構造を内包する求人票	40

表一覧

2.1	求人票の種別ごとの件数（白地以外の背景のある求人票を含む）	9
2.2	求人票の種別ごとの件数（白地以外の背景のある求人票を含まない） . .	9
3.1	住居表示（建造物名など）に出現する単語と頻度	21
4.1	内容の抽出率に関する実験結果	30
4.2	求人票中の内容を企業データベースへ手入力するのに要する時間（求人票 10 枚あたり）	36
4.3	提案法による企業データベースの作成時間（求人票 10 枚あたり）	37
4.4	誤読文字を修正せずに文字認識結果を作成する時間（求人票 10 枚あたり）	38

第1章

はじめに

1.1 研究の背景

現在，2014年3月卒業予定（以下，2014年卒大学生）の大学生による就職活動が本格化している。就職活動では，一般的にリクナビやマイナビなどの全国規模の就職支援システム [1][2][3][4] が利用されている。これらのシステムでは，全国各地の学生が全国各地の企業を調査することができ，入社を希望する企業の説明会や面接等を予約する機能が備わっている。しかし，説明会や面接会場までの距離を理由に，実施日の直前に予約を取り消す学生が多く，企業では多大な事務処理が発生している。特に中小企業では，採用人数が多い時で数名である企業が多く，人事担当者は少ない。そのため，これらの事務処理や求人システムへの毎年数百万円に及ぶ登録・掲載費用が大きな負担となり，採用活動の妨げとなっている。

近年の就職活動における特徴の1つに学生の「大手志向」がある。文献 [5] によると，2013年3月卒大学生を対象とした従業員規模別の求人倍率（＝求人総数/民間企業就職希望者数）では，従業員1000人未満企業で1.79倍，従業員1000人以上企業で0.73倍となっている。このように，依然として学生の目は大企業に向けられている。しかし一方で，2010年以降での従業員規模別の就職希望者数の推移をみると，学生が年々，中小企業への関心を強めている傾向が伺える [5]。そのため，中小企業にも焦点を当てた就職支援の必要性が増加している。

中小企業に焦点を当てた就職支援システムの一例として，三重県四日市商工会議所では，三重県の企業を対象とした就職支援システムである「三重就職NAVI」[6]を運用している。三重就職NAVIでは，勤務地が三重県の中小企業に対して主軸を置き，新卒採用や中途採用の求人を行なうことにより，中小企業の採用活動を支援している。

また，2013年卒大学生を対象とした就職意識調査によると，学生が企業選択において最も重視する基準は「自分のやりたい仕事（職種）ができる会社」であり，次いで「働

きがいのある会社」と続く [7]。しかし、学生は業界や企業に関する知識が乏しく、それらの情報を得る機会が少ないのが現状である。そのため、学生は企業の業務内容を理解することができず、仕事の魅力、企業の魅力、将来の展望を明確にできていない。その結果、仕事の魅力、企業の魅力、将来の展望を軸とした就職活動が困難となり、学生はテレビでコマーシャルを流すような有名企業への就職を希望する傾向がある。

以上のことから、学生が「やりたい仕事ができる会社」、「働きがいのある会社」を就職先として選択するために、企業の業務内容を学生に理解しやすく提示することができ、企業の採用活動の負担を大幅に軽減する就職支援システムの必要性が高まっている。

1.2 提案システムの概要と本論文の位置づけ

本研究では、学生が企業の業務内容をイメージする指標として、大学の授業科目に着目し、授業科目と関係性の高い企業等の就職情報を紹介する「就職相談システム」の開発を目的とする。まず、三重大大学の電気電子工学科を対象とした就職支援システムの開発をめざす。このシステムでは、授業科目を利用する観点から、特定の大学の特定の学科に特化し、企業データベース情報には、学科に送付される求人票を利用する。現状の三重大大学では、紙形式の求人票や会社案内を箱に並べ、手書きでリスト化して保存し、学生が閲覧可能な状態にしており、利便性が低い。そこで本システムでは、求人票や学科に蓄積された就職情報（卒業生の就職情報、職場での体験談等）を電子化して、Web上で閲覧できる仕組みを採用し、学生や企業への情報提供を効率的かつより適切に行うことができる。

将来的には、学生が授業科目、希望職種等を入力すると、システムが企業情報や授業科目に関するデータベースを検索することにより、関係性の高い企業の求人票や業界マップ、卒業生の就職情報、職場での体験談等を出力するシステムを想定している（図 1.1）本研究では、まず三重大大学の電気電子工学科を対象とした就職支援システムの開発をめざす。

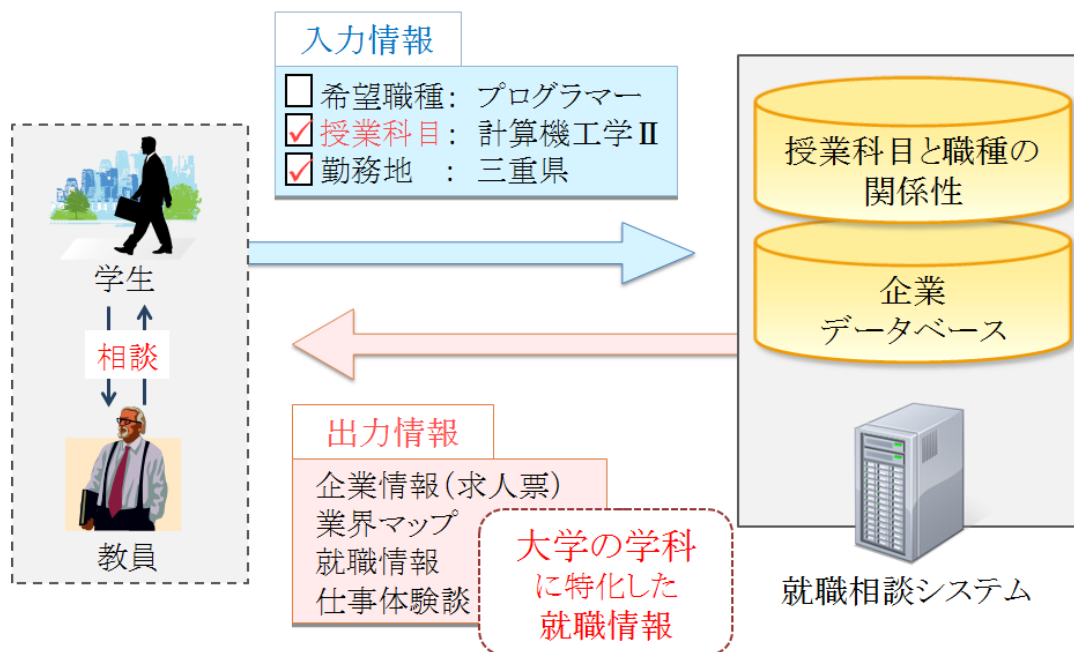


図 1.1: 就職相談システムの使用イメージ

システムでは、企業情報を登録した企業データベースに加えて、授業科目と職種の関係性を示すデータベース（授業科目と職種の関係データベース）を作成して用いる。その開発過程は、授業科目と職種の関係データベースを作成する過程と企業データベースを作成する過程、システムのユーザーインターフェイスを開発する過程からなる（図1.2）。筆者の在籍する研究室では、授業科目と職種の関係性の導出とシステムのユーザーインターフェイスの開発に関する研究を行ってきた [8]。授業科目と職種の関係データベースを作成する過程では、学科の授業科目のリストを作成し、企業データベースに格納された職種データを利用する。研究では、これらの授業科目と職種データから授業科目と職種の関係性を示す表（図1.3）を作成し、それらの関係性の導出を試みている。表の評価値を反映させた「授業科目と職種の関係データベース」を作成し、企業データベースと対応付けることで授業科目から企業情報の検索を可能にしている。

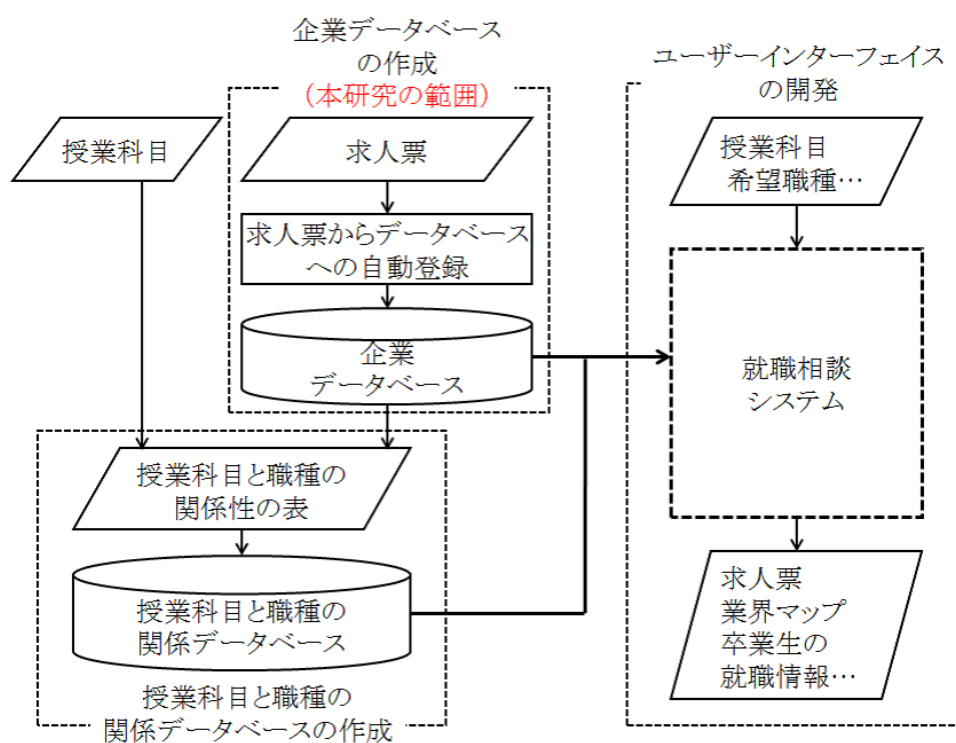


図 1.2: 就職相談システムの開発過程

		メーカー										商社	百貨店・ ストア・専 門店	金融・証 券・保険	情報(通信・ マスコミ)	ソフトウェ ア・情報 処理
		建設	食品	繊維・紙	化学・ 薬品	石油・ 窯業	鉄鋼・ 金属	機械	電機	輸送機器	精密・諸 工業					
材料	材料	1	1	2	2	2	2	2	2	2	1	0	0	0	0	
	物性	1	1	2	2	2	2	2	2	2	1	0	0	0	0	
電機	電力	1	1	1	2	2	2	2	2	2	1	0	0	0	0	
	電機	1	1	1	2	2	2	2	2	2	1	0	0	0	0	
情報	情報	1	1	1	2	1	1	2	2	2	1	1	2	2	2	
	通信	0	0	0	0	0	0	1	2	2	1	0	0	2	2	
その他	外国語	1	1	1	1	1	1	1	2	2	1	2	2	2	2	
	数学	1	1	1	1	1	1	2	2	2	1	1	2	2	2	
	物理	2	1	1	1	1	2	2	2	2	1	1	1	1	1	

図 1.3: 授業科目と職種の関係性の表（一部）

1.3 現在の問題点と従来研究の概要

こうした大学で運用可能な就職支援システムの開発は盛んに行われている [9][10][11][12] . 中でも, 求人情報の閲覧システムに関する研究では, その多くがシステムの出力ページの情報ソースとして求人票を利用している. しかし, 単一の学科に送付される求人票の数は数百～数千件 (本学電気電子工学科における平成 24 年の求人票総数は約 500 件であ

る)にも及び、就職事務担当者の手作業によるシステムの運用は困難である。そのため、様々な過程の自動化に関する研究が進められている。例えば求人票を対象とした、求人情報の抽出及び閲覧ページへの自動出力に関する試みがある。

文献[9]では、あらかじめ指定した Excel の求人マスタファイルに対して、就職事務担当者が求人票中の相当する内容をテキスト入力し、このファイル情報を出力とした求人情報の閲覧ページを自動作成している。また文献[10]では、求人情報の閲覧ページの出力として求人票の画像データを利用し、就職事務担当者が画像データのファイル名に「会社名、業種、地域名、ページ番号を入力し、このファイル名を基に求人票の閲覧ページを自動作成している。これらの研究では、就職事務担当者が求人情報の閲覧ページを作成する手間を削減しているが、閲覧ページの自動作成に使用する求人情報は依然として就職事務担当者の手入力であり、多大な手間を要する。

本研究では、会社名、業種を始めとする多種の求人情報を情報検索で利用するため、これらを企業データベースへ格納する必要がある。求人票中の情報を手作業でデータベースに入力する手間と時間は更に大きい。そのため、例えば三重大学では、現在の事務組織の人員によるシステムの運用が困難である。したがって、求人票の内容を自動的に抽出し、企業データベースへ適切に格納するアルゴリズムを開発する必要がある。求人票などの表記形式が多様な文書を対象とし、文字認識して項目ごとに分類する市販ソフト[13][14]や研究[15][16]がこれまで進められてきた。

文献[15]では、名刺に表記される人名、組織名、住所及び電話番号に注目し、これらを分類して抽出している。項目を特定するため、固有名詞のデータベースとの比較を行う。具体的には、人名の特定には姓名データベース、組織名の特定には組織名からなる固有名詞データベースを利用している。また文献[16]では、名刺に表記される氏名、組織名、所属、所在地などに注目し、これらを分類して抽出している。項目を特定するため、文字認識誤りを考慮したキーワード辞書とのロバストキーワードマッチングを行い、予め用意した文書モデルと連合グラフ法に基づくモデルマッチングを行う。本研究では、求人票を対象とし、項目数、罫線の表記方法や文字列の配置方法が名刺に比べて遥かに多様である。そのため、文献[15]の手法のみでは、高精度での項目の分類が見込めない。また、求人票の項目数、罫線の表記方法や文字列の配置方法を企業により大きく異なるため、文献[16]の手法では、利用する文書モデルが求人票では膨大となるため、システム作成の実現性と処理時間の観点から現実的ではない。

1.4 本研究の目的及び概要

本研究では、就職相談システムで利用する企業データベースを作成するため、求人票の内容を抽出してデータベースへ自動登録する手法を検討する。求人票の表記形式は多種多様であるため、定型書式を文書認識する市販の文字認識ソフトウェアを使用して、求人票の文字列をデータベースへ格納することは困難である。そこで本論文では、まず活字で表記された紙形式の文書もしくはその画像データ(jpg)を対象として、市販の文字認識ソフトウェアにより表形式の文字認識を行う。その文字認識結果中の文字列を「項目名」と「内容」ごとに整理し、整理した文字列(統一フォーマット)からキーワード(現状では、会社名、本社所在地、職種、勤務地、福利厚生、資本金、初任給、従業員数)のみをデータベースへ格納する手法を提案する。

本研究により多様な文書形式を処理し、企業データベースへの自動登録システムを構築する研究を進めれば、より一般的な処理システムを開発する場合の「多様性を吸収する処理モジュール」と「共通処理モジュール」によるモジュール分割方法などのシステム開発方法を体系的に考察することが可能となる(図1.4)。「多様性を吸収する処理モジュール」と「共通処理モジュール」に分割したシステム開発により、処理対象の多様性、つまり処理対象ごとの異なる長所を活かしたシステムの開発ができる。

例えば、図1.2のシステムを完成させる場合では、企業データベースへの自動登録処理過程は、紙ベースの多様な求人票についての画像入力による処理なので「多様性を吸収する処理モジュール」であり、その他の過程は、キーボードによる入力なので、データを統一形式で入力する処理となり、「共通処理モジュール」となる。求人票では、文書ごとに表記される項目と表記されない項目が異なり、それぞれの項目に有用性があるため、それらの多様性を企業データベースへ反映させることは重要である。

本論文では、全ての項目名と内容が種別ごとに罫線で区切られている求人票を対象としている。また学生が就職活動をする際に、企業情報として重要視する「会社名と本社所在地」に焦点を当て、提案法の有効性や精度に関する評価実験を行うと共に考察を行う。さらに、現状の問題点について考察し、今後の課題点についても述べる。

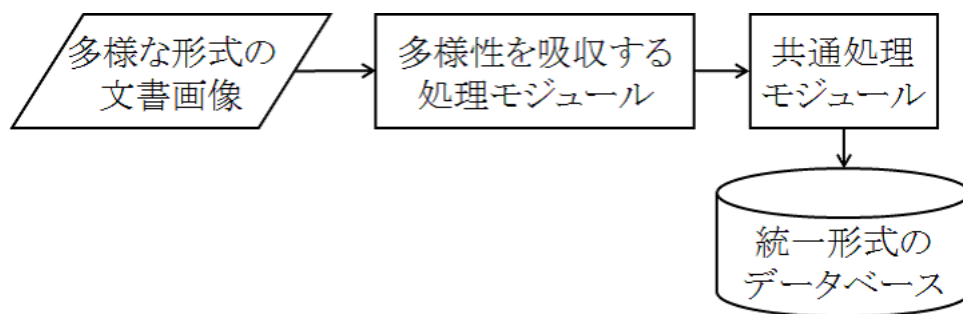


図 1.4: 多様性の吸収モジュール

第2章

研究対象とする文書画像

2.1 求人票

求人票とは、職業安定法によって定められた労働条件を明示した書類である [17]。具体的には、以下の項目が明記されている。

1. 労働者が従事すべき業務の内容に関する事項
2. 労働契約の期間に関する事項
3. 就業の場所に関する事項
4. 始業及び終業の時刻、所定労働時間を超える労働の有無、休憩時間及び休日に関する事項
5. 賃金の額に関する事項
6. 健康保険、厚生年金、労災保険、雇用保険の適用に関する事項

就職希望者は、求人票を熟読することにより、企業の労働条件を把握することができる。新卒大学生を募集する企業では、求人票を採用予定のある特定の大学の特定の学科に送付することが多い。以上の理由から、本研究では学科単位で取り扱う求人情報として、求人票を使用することとした。

2.2 使用する求人票

本学電気電子工学科の学生を対象として送付された、2006 年卒大学生向けの求人票は 234 枚であり、2013 年卒大学生向けの求人票は 501 枚である。これらの求人票は、紙媒体であり活字で表記されている。本研究では、求人票を画像データ（解像度 400dpi、圧縮形式 jpeg）へ変換して使用する。

本研究では、求人票の文書画像を文字認識して利用する。求人票の表記形式は多種多様であり、文字認識結果において多様性を示す要素では「罫線の割合」、「白地の背景が

どうか」が挙げられる．特に白地以外の背景を含む求人票では，白地以外の背景部分に記入された文字列を文字認識できない場合があり，それにより文字認識結果を対象とした後述する処理が適用できない可能性がある．

そこで，2013年卒大学生向けの求人票を，「全体が罫線による表構造をなす求人票（図2.1）」，「一部が罫線による表構造をなす求人票（図2.2）」，「文字のみによる求人票（図2.3）」，「白地以外の背景のある求人票（図2.4）」に分類した．

ここでの全体が罫線による表構造をなす求人票とは，全ての項目名と内容を種別ごとに罫線で区切る構造を採用している求人票を指す．また，一部が罫線による表構造をなす求人票とは，一部の項目名と内容を罫線で区切る構造を採用している求人票を指す．

これらの求人票の件数と総数に占める割合を表2.1，表2.2に示す．表2.1，表2.2では，全体が罫線による表構造をなす求人票が，いずれも求人票全体の過半数を占めている．そのため本稿では，全体が罫線による表構造をなす求人票に焦点を当てている．

表 2.1: 求人票の種別ごとの件数（白地以外の背景のある求人票を含む）

求人票の構造		求人票数 [件]	占有率 [%]
白地の背景	全体が罫線による表構造	274	54.7
	一部が罫線による表構造	43	8.6
	文字のみ	48	9.6
白地以外の背景		136	27.1
計		501	100.0

表 2.2: 求人票の種別ごとの件数（白地以外の背景のある求人票を含まない）

構造	求人票数 [件]	占有率 [%]
全体が罫線による表構造	376	75.0
一部が罫線による表構造	60	12.0
文字のみ	63	12.6
計	501	100.0

会社名	株式会社
所在地	〒510-8521 三重県三重郡朝日町縄生2121
代表者	取締役社長
設立	1998年1月
資本金	490百万円 (出資:(株) 49%)
売上高	118億円 (2003年度)
従業員数	122名 (2004年1月現在)
事業内容	産業用インバータの開発・製造
事業場	本社
求人職種	研究開発、設計
勤務時間	8:00~16:45 標準労働時間7時間45分 フレックスタイム制あり
休日・休暇	完全週休2日制 有給休暇:初年度18日、2年目以降24日


図 2.1: 全体が罫線による表構造

2013 年度大卒募集要項

23.12.27

株式会社 製作所

78

<p><募 集 要 項></p> <ul style="list-style-type: none"> ・採用学科 理 系：学部学科不問（機械工学系積極採用） ・採用職種 設計開発、生産管理、セールスエンジニア ・採用予定 12 名程度 ・勤 務 地 鴻池事業所（東大阪市東鴻池町 2-1-48） ・初 任 給 修士課程修了 月額 214,700 円 学部卒業 月額 200,700 円 ・諸 手 当 営業／設計開発手当（月額 32,000 円）、家族手当、 役職手当、通勤手当（実費） ・昇 給 年 1 回（4 月）5,300 円 (2011 年度組員平均の実績) ・賞 与 年 2 回（6 月、12 月）年間実績 150 万円＋利益配分 (2011 年度組員平均) ・勤務時間 8:30～17:00（支店 9:00～17:30） ・休日休暇 完全週休 2 日制、年次有給休暇初年度 12 日、年間 休日 124 日、慶弔特別休暇、連続休暇（夏季、 年末年始、ゴールデンウィーク） ・福利厚生 各種社会保険、財形貯蓄、持株会、住宅ローン補助、 福祉会、OB会、クラブ活動、レクリエーション（補 助）、ワンルームマンション斡旋、 家賃補助 45,000 円（対象は単身者で通勤圏外の方） ・教育制度 新入社員集合研修、職能別専門教育、OJT、 自己研修、通信教育、外部研修 ・提出書類 履歴書（写真貼付）、成績証明書（含卒業見込証明 書）、健康診断証明書 ・選考方法 面接、健康診断、適性検査 	<p><会 社 概 要></p> <ul style="list-style-type: none"> ・創 立 昭和 17 年 5 月 ・業 種 熱交換器、食品・化学機械、モノのモノ、染色仕上機械、 産業機械の製造販売 ・資 本 金 41 億 5000 万円 ・株式上場 東京、大阪証券取引所（市場第一部） ・年 商 208.4 億円 ・従 業 員 450 名（男 396 名、女 54 名、平均年齢 36.1 才） ・役 員 代表取締役会長 代表取締役社長 専務取締役 常務取締役 常務取締役 取 締 役 取 締 役 取 締 役 取 締 役 <p><採用に関する連絡先></p> <p>本社／大阪市中央区伏見町 4 丁目 2 番 1 4 号</p> <p>人間課 TEL (06) 6201-3531 鴻池事業所／東大阪市東鴻池町 2 丁目 1 番 48 号 人間課 TEL (072) 966-9600 東京支店／東京都中央区京橋 1 丁目 1 1 番 2 号 人間課 TEL (03) 5250-0750 URL http://www.hisaka.co.jp E-mail jini@hisaka.co.jp</p>  <p>携帯QRコード：リクナビ2013</p>
--	--

最近 5 年間の実績

年 度	2006 年	2007 年	2008 年	2009 年	2010 年
売上高 億円	248.9	291.9	350.9	244.7	208.4
経常利益 億円	43.1	50.7	50.1	18.2	14.6
経常利益率 (%)	17.3	17.4	14.3	7.4	7.0
自己資本比率 (%)	76.6	68.5	72.4	87.8	84.4

売上構成比

熱交換器	生活産業機器	バルブ
56.8%	30.0%	13.2%

図 2.2: 一部が罫線による表構造

株式会社 **求人票**
(2013年3月卒業見込者及び2010年3月以降の既卒業者対象)

●会社概要

会 社 名 株式会社 **Hitachi Information & Control Solutions, Ltd**
 茨城本社 茨城県日立市大みか町5-1-26
 東京本社 東京都台東区秋葉原6-1 (秋葉原大栄ビル)
 代 表 者 取締役社長 **取締役社長**
 設 立 2006年4月
 資 本 金 22億7千万円
 社 員 数 2,759名
 売 上 高 479億円(2011年3月期)
 事業内容 【システムインテグレーション】
 ・システムコンサルタント
 ・システム開発
 ・システム保守
 【ソリューションサービス】
 ・システム構築
 ・アプリケーションパッケージソリューション
 ・オープンソリューション
 【ハードウェア・ソフトウェア開発】
 【情報機器販売】
 認定認証 CMMIレベル5達成、ISO9001認証取得、
 ISO14001認証取得、プライバシーマーク取得、
 特定建設業(電気工事業、電気通信工事業)
 特定労働者派遣事業
 関連会社 茨城日立情報サービス株式会社

休日・休暇 完全週休2日、春季・夏季・年末年始長期休暇、
 年次有給休暇(22~24日、計画年休、半日年休制度
 有り)、リフレッシュ休暇制度、年間休日約130日
 福利厚生 保険／ 各種社会保険、年金制度、財形貯蓄
 施設／ 従業員クラブ、スポーツ施設、
 病院、リゾート施設等有
 住居／ 独身寮、社宅、住宅手当制度有り

●募集要項

募集職種 システムエンジニア(システムコンサルタント)
 ソフトウェアエンジニア(ソフトウェア開発・設計)
 ハードウェアエンジニア(ハードウェア開発・設計)
 セールスエンジニア(営業)
 経営サポートスタッフ(管理部門)
 採用学科 全学部全学科
 応募方法 学校推薦 又は 自由応募
 提出書類 履歴書、成績証明書、卒業見込証明書、推薦書
 ※成績証明書、卒業見込証明書の発行が開始されていない
 場合は、ご相談下さい。
 ※学校推薦の場合は、推薦書も併せてご提出いただきます。
 書類提出先 下記の問い合わせ先の担当宛に送付願います。
 選考方法 書類選考(推薦応募は免除)、適性検査、面接
 選考日時 詳細は本人へ通知致します。
 選考時携行品 筆記用具、印鑑

●待遇・勤務条件

初 任 給 修士了 228,500円(2011年4月実績)
 大学卒 205,500円(")
 高専卒 180,500円(")
 給 与 改 定 年1回(4月)
 賞 与 年2回(6月、12月)
 通 勤 費 全額支給
 勤 務 地 東京地区(東京都台東区)
 茨城地区(茨城県日立市)
 勤務時間 8:50~17:20(東京地区)
 8:40~17:10(茨城地区)
 フレックスタイム制度有り

●問い合わせ先

〒110-0006
 東京都台東区秋葉原6-1 (秋葉原大栄ビル)
 株式会社 **日立情報サービス株式会社**
 総務部 採用担当 **採用担当**
 電 話: 0120-527-440 (採用フリーダイヤル)
 電 話: 03-3251-7620 (直通)
 FAX: 03-3251-7210
 E-mail: saiyu@

図 2.3: 文字のみによる求人票

株式会社	
<p>「温もりある明日のために。」</p> <p>「Paloma」。スペイン語で「鳩」という意味です。 世の中に平和で豊かな暮らしを。 1911年の創業以来、そんな想いを込めて100年以上にわたり事業を展開してきました。 この想いは海を越え、アメリカ・アジア・オーストラリアなど拠点をもち、60カ国で販売実績を残すまでに。勢いは現在進行形で増しており、特に近年、ガスエネルギーへの期待が追い風になっています。 世間から寄せられる一つひとつの意見に耳を傾け、製品に反映させる。いつの時代も、地に足をつけた事業展開をしています。 これも、生活に密着した製品を扱っているからこそ。 この姿勢は、これからも変わりません。 日本、そして海外各国に、平和で豊かな生活を。 私たちの挑戦は続きます。</p>	
<p>企業概要(グループ)</p> <p>企業名 株式会社</p> <p>創業 1911年2月 代表者</p> <p>本社所在地 〒467-8585 愛知県名古屋瑞穂区桃園町6番23号</p> <p>売上高 2,500億円(2011年2月連結実績)</p> <p>従業員数 14,928名(グループ計)</p> <p>事業内容 【家庭用・業務用ガス器具の開発・製造・販売】 ・厨房機器／ガステーブルコンロ、ビルトインコンロ、炊飯器 ・温水機器／ガス給湯器、床暖房システム、浴室乾燥機 ・業務用機器／フライヤー、ロードヒーティング、ガス炊飯器 ・次世代エネルギー／太陽熱温水器、燃料電池</p> <p>本社(名古屋市)</p>	
主要事業所	<p>工場 本社工場・本社第2工場・清洲工場・大口工場(愛知県)、北勢工場・笹野工場(三重県)、恵那工場・可児工場(岐阜県)、直方工場(福岡県)、北海道工場(北海道登別市) 他</p> <p>営業所 アラバマ・アーカンソー・カリフォルニア・フロリダ(以上アメリカ)、シドニー・メルボルン・パース(以上オーストラリア)、成都(中国) 他</p> <p>海外事業所 札幌・仙台・東京・新潟・金沢・静岡・名古屋・大阪・広島・高松・福岡・沖縄・その他全国各地 合計78ヶ所</p> <p>研究所 本社・札幌</p>
ホームページ	http://www.paloma.co.jp
情報掲載サイト	【リクナビ2013】にて採用情報を掲載しております
募集要項	
募集職種(総合職)	<p>技術職…製品開発、要素技術開発、電子技術開発、生産技術、製造エンジニア 他</p> <p>営業職…国内営業、海外営業</p> <p>本社スタッフ…人事部、経理部、原価企画部、情報システム部、生産管理部、品質管理部 他</p>
採用予定人数	<p>大学学部卒業予定者 } 技術職20名、営業職20名、本社スタッフ若干名</p> <p>大学院修士課程・博士課程修了見込者 }</p>
初任給	大学卒…210,000円 大学院修了…228,000円 ※2011年実績
昇給	年1回(4月) 賞与 年2回(7月・12月) 平均4.0ヶ月 ※2011年実績
休日休暇	当社カレンダーによる(原則土日祝休み) その他夏季・GW・年末年始 年間休日計116日
採用試験	
応募形式	自由応募、学校推薦(一部対象校)
応募方法	<p>【リクナビ2013】よりエントリーして頂き、説明会へのご予約・ご参加をお願いします</p> <p>◆説明会にてエントリーシートをお渡しします</p> <p>◆説明会は2012年2月中頃より、札幌・東京・名古屋(本社)・大阪・福岡にて順次開催します</p>
選考方法	エントリーシート、筆記試験、適性検査、面接試験
提出書類	<p>1.履歴書(写真添付) 2.成績証明書 3.卒業見込証明書 4.健康診断書</p> <p>※全て最終面接時に提出</p>
お問合せ先	
<p>〒467-8585 愛知県名古屋瑞穂区桃園町6-23</p> <p>人事部 (採用担当)</p> <p>TEL: 052-824-5167</p>	

図 2.4: 白地以外の背景のある求人票 (画像はグレースケール)

2.3 求人票の文字認識結果

求人票は表形式の基本構造を有し，記入された項目名と内容が左右で隣接し，対応関係を示す．そのため，一般的に求人票の文字認識結果では，項目名とその内容の関係を示すデータが連続して表記されている．本手法では，求人票の画像データに対して市販の文字認識ソフト（表 OCR / 文書 OCR for Excel & Word，富士通ミドルウェア株式会社製）を用いて文字認識を行う [18]．

図 2.5 に，求人票とその文字認識結果の例を示す．図より，文字認識結果では以下の特徴が表れている．

1. 求人票の表罫線がコンマとして表示される．
2. スペース，改行，ダブルコーテーションや全角文字と半角文字が混在している．

使用した文字認識ソフトでは，求人票中の罫線，空白，行末が，それぞれコンマ，スペース，改行に変換される．これらの変換機号は OCR ソフトウェア毎に異なるが，変換アルゴリズムはほぼ同じであり，記号を変更することで容易に対応できる．また，市販の文字認識ソフトにより作成した，求人票の文字認識結果では，表構造や文字を再現できていない場合がある．誤認識した文字は，提案する一連の処理に対して重大な悪影響を与えるだけでなく，就職相談システムの企業データベース内容としても望ましくない．そこで，市販の文字認識ソフトに対して一般的に実装されている「文字の修正機能」を利用し，手作業で修正したものを後述の処理で利用する．なお，誤認識した表構造については，市販の文字認識ソフトの標準機能を利用した修正が困難であるため，本研究では，大幅な修正を加えていない．本研究では，まず市販の文字認識ソフトを使用して求人票を文字認識する場合の問題点を明確にし，その解決策について検討する．将来的には，市販の文字認識ソフトを使用した場合における問題点の解決策を反映させた，求人票に特化した文字認識ソフトを作成する．それにより，この過程をある程度自動化し，システム運用担当者の手間と作業時間をより削減できるシステムをめざす．

第3章

求人票画像から企業データベースへの自動登録

3.1 企業データベースの概要

本研究における企業データベースとは、提案システムの入力情報及び出力情報の内、企業に関する情報を格納したデータベースを指す。具体的には、会社名、本社所在地、業種、職種、初任給、資本金、卒業生の就職情報、職場での体験談等を格納する。

図3.1は、企業データベースのテーブル構造であり、求人票から抽出したデータのフォーマットはこのテーブル構造に準ずるものとした。また、図3.2に企業データベースへの自動登録手順を示す。

企業ID	企業名	所在地名	勤務地	初任給(大学)	初任給(大学院)	職種ID	職種名
		〒140-0013東京都品川区南大	東京	201,000	222500	2	海外営業
		〒140-0013東京都品川区南大	東京	201,000	222500	14	システムエンジニア
		〒140-0013東京都品川区南大	東京	201,000	222500	15	ネットワークエンジニア
		〒140-0013東京都品川区南大	東京	201,000	222500	16	ソフトウェア開発
		〒140-0013東京都品川区南大	東京	201,000	222500	17	システム運用・保守
		〒460-0008名古屋市中区栄2-	愛知	201,000	222500	20	商品開発・設計
		〒113-0021東京都文京区本駒	東京	201,000	222500	14	システムエンジニア
		〒113-0021東京都文京区本駒	東京	201,000	222500	15	ネットワークエンジニア
		〒113-0021東京都文京区本駒	東京	201,000	222500	21	生産・製造技術開発
		〒485-8551愛知県小牧市応時	愛知	200,300	214300	13	セールスエンジニア
		〒485-8551愛知県小牧市応時	愛知	200,300	214300	13	セールスエンジニア
		〒485-8551愛知県小牧市応時	愛知	200,300	214300	13	セールスエンジニア
		〒485-8551愛知県小牧市応時	愛知	200,300	214300	14	システムエンジニア
		〒485-8551愛知県小牧市応時	愛知	200,300	214300	15	ネットワークエンジニア
		〒485-8551愛知県小牧市応時	愛知	200,300	214300	16	ソフトウェア開発
		〒100-0005東京都千代田区丸	東京	200,000	225500	18	基礎研究
		〒222-8540神奈川県横浜市港	神奈川	201,000	222500	15	ネットワークエンジニア
		〒222-8540神奈川県横浜市港	神奈川	201,000	222500	16	ソフトウェア開発
		〒105-8659東京都港区新橋1-		201,000	222000	17	システム運用・保守
		〒105-8659東京都港区新橋1-		201,000	222000	18	基礎研究
		〒105-8659東京都港区新橋1-		201,000	222000	19	応用研究・技術開発
		〒105-8659東京都港区新橋1-		201,000	222000	20	商品開発・設計
		〒105-8659東京都港区新橋1-		201,000	222000	21	生産・製造技術開発
		〒105-8659東京都港区新橋1-		201,000	222000	22	生産・工程管理

図 3.1: 企業データベースのテーブル構造

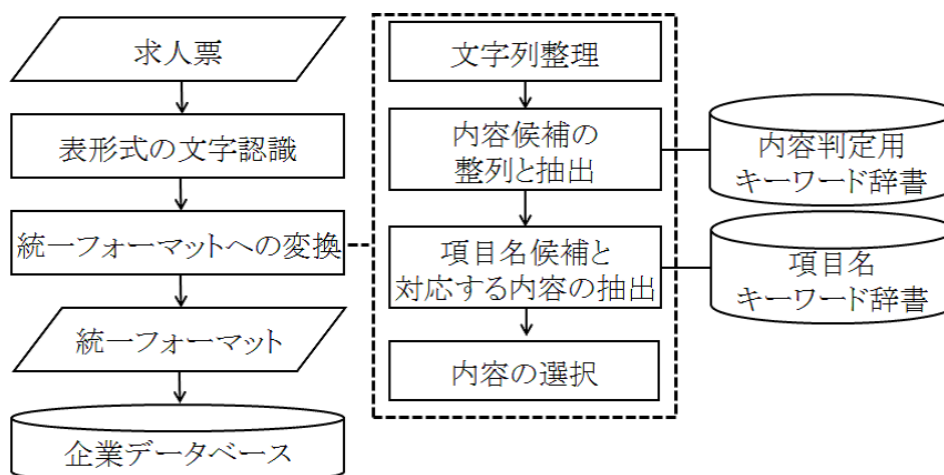


図 3.2: 企業データベースへの自動登録手順

3.2 統一フォーマットへの変換

求人票の表記形式が多様多様のため、文字認識結果の表記形式も多種多様となる。そのため、文字認識結果の文字列を一様にデータベースへ格納することは困難である。そこで本手法では、図 3.3 に示すように、求人票の文字列を「項目名」と「内容」ごとに整理し、データベースへ格納する。

求人票から目的の内容を抽出するため、人が求人票から目的の内容を特定する方法に注目した。以下に、方法をまとめる。

方法

1. 目的の内容に対応する項目名を特定し、項目名の付近にある内容を特定する。
2. 目的の内容を示す文字列が含むと予想される単語を探し、目的の内容を特定する。

以上を受けて、求人票から目的の内容を抽出するための方針を大きく 2 つに定めた。以下に、方針をまとめる。

方針

1. 項目名判定用キーワード辞書を用いて、項目名の候補となる文字列を特定し、項目名候補と対応する内容の位置関係から内容を決定する。
2. 内容判定用キーワード辞書を用いて、内容の候補となる文字列を決定する。

方針 1, 2 により抽出した項目名候補と内容候補による文字列から、企業データベースへ登録する内容を選択し、「統一フォーマット」を作成する。企業データベースでは、CSV ファイルによるデータの入力が可能である。そこで、求人票から抽出した項目名と内容に対して文字列のフォーマットを、3.1 節で述べた企業データベースのテーブル構造に準

ずる形で、項目名と内容ごとに2行形式で表記した文字列（CSV形式）に変換する。本研究では、CSV形式に変換した文字列を統一フォーマットと定義する。

本節では、文字認識結果の文字列を統一フォーマットへ変換する方法について述べる。なお現段階では、求人票の重要項目として、「会社名」と「本社所在地」に焦点を当てている。

項目名	会社名, 本社所在地, 勤務地, 初任給…
内容	〇〇株式会社, △△県～市～, ××県, ～～…

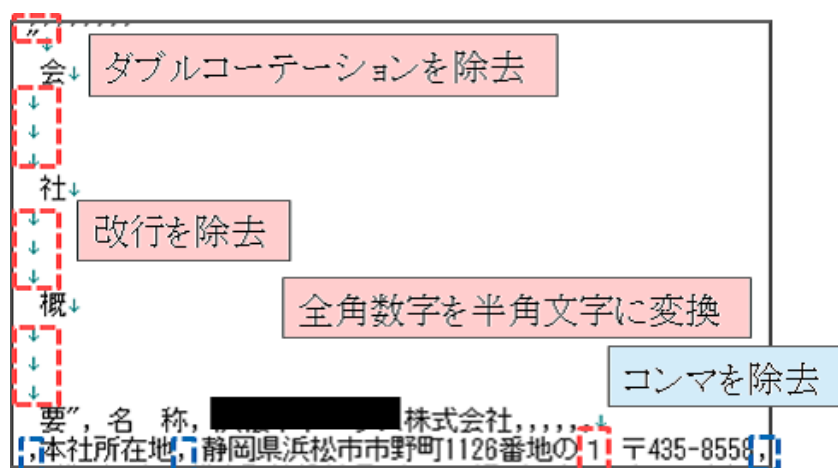
図 3.3: 統一フォーマットの例（2行のCSV形式）

3.2.1 文字列整理

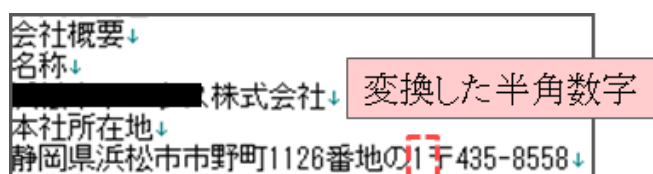
求人票は共通的な特徴として、左右のデータ間で対応関係を示す表形式の基本構造を有する。そのため求人票の文字認識結果は、項目名と内容の対応関係を示すデータが連続して表記されている。この特徴により連続するデータ間を切り分けることで、項目名と内容ごとに整列させることが可能である。そこで、2.3節で述べた文字認識結果の特徴である、「求人票の罫線を示すコンマ」を利用して、データごとに分離する。また文字認識結果では、さまざまな字種が混在しており、出力形式の統一化と後述する処理のため、これらの文字列の整理を行う。以下に、具体的な処理の流れを示す。

1. コンマ間の文字列を分離する。
2. 全角英数字及び全角カッコを半角文字へ変換し、半角カタカナを全角文字へ変換する。
3. 整理結果を「文字列整理処理の途中結果」として保存する。
4. スペース、改行、ダブルコーテーションを除去する。

図 3.4(b) に、図 3.4(a) の文字認識結果に対して文字列整理処理を適用した後の文字列の状態を示す。図では、「会社概要」が同一行に格納され、ダブルコーテーション、改行記号及びスペース記号が除去され、全角文字の「1（数字のイチ）」が半角文字に統一されている。すなわち、文字列が整理され、項目名と内容ごとに分離して整列されていることがわかる。



(a) 文字認識結果



(b) 文字列整理処理の適用結果

図 3.4: 文字認識結果（上）に対する文字列整理処理の適用結果（下）

3.2.2 内容候補の整理と抽出

3.2 節で述べた方針 2 に定めた内容候補の判定を行う。

方針 2 では、内容を表す文字列を特定する必要がある。まず、求人票 140 件から「会社名」と「本社所在地」に関して、内容の表記方法を調査した。すると、会社名には必ず企業形態（株式会社など）が付随し、本社所在地には（求人票 140 件中 136 件で）住所表記が記入されていることがわかった。そこで、企業形態や住所表記における区切り文字、都道府県名を登録した単語辞書（以下、内容判定用キーワード辞書）を独自に作成した。図 3.5 に、単語辞書の中身を示す。

また特に住所表記の判定では、表記の法則性に注目し、財団法人地方自治情報センター [19] が作成する「住所コード」から、独自に住所の階層構造モデルを作成した。図 3.6 に、住所の階層構造モデルを示す。しかし、住所コードでは住居表示（建造物名など）に用いられる単語に関して詳しく言及していないため、独自に求人票 85 件における住居表示を調査し、建造物名の表記方法を単語辞書へ反映させた。なお、住所表記中に住居表示が出現したのは求人票 140 件中 48 件だった。出現した単語とその頻度を表 3.1 に示す。

内容判定用キーワード辞書と住所の階層構造モデルを利用した、内容候補の判定方法（方針 2）を以下に示す。内容候補の判定（方針 2）で求めた内容候補の記入位置は、後

述する「1つの項目名に対して、対応する複数の内容が表記されている求人票」や「内容のみが表記され、対応する項目名が表記されていない求人票」などの、多様性を示す求人票を統一フォーマットへ変換する場合に、内容候補を示す文字列を特定するために利用する。

1. 内容判定用キーワード辞書と求人票の文字列整理結果間で、単語の一致判定を行う。

- 会社名に関する内容の場合

2. 一致した単語を含む文字列を、会社名の内容候補として判定し、その記入位置を特定する。

- 本社所在地に関する内容の場合

2. 一致した単語により、使用する住所の階層構造モデルを決定する。

3. 一致した単語を含む文字列で、一致した単語の出現順序が住所の階層構造モデルに従うかどうか判定する。

4. 住所の階層構造モデルに従い、一致した単語の出現数が4以上のとき、本社所在地の内容候補として判定し、その記入位置を特定する。

5. 判定された内容候補がないとき、一致した単語の出現数が3以上にして、内容候補の記入位置を特定する。

なお、単語の出現数を4としたのは、予備実験による経験的な値による。

表 3.1: 住居表示（建造物名など）に出現する単語と頻度

住居表示	出現数 [件]
ビル	35
館	7
タワー	8
階	10
F(Floor)	8
計	68

〈社名判別〉	愛知県	〈住所判別〉	0	番号
株式会社	三重県	郡	1	の
株式會社	滋賀県	市	2	番地
(株)	京都府	町村	3	ビル
(株)	大阪府	字	4	館
	兵庫県		5	階
〈都道府県判別〉	奈良県	0	6	F
北海道	和歌山県	1	7	タワー
青森県	鳥取県	2	8	階
岩手県	島根県	3	9	F
宮城県	岡山県	4	番号	の
秋田県	広島県	5	番号	番地
山形県	山口県	6	の	区
福島県	徳島県	7	の	町村
茨城県	香川県	8	番号	丁
栃木県	愛媛県	9	の	0
群馬県	高知県	番号	番地	1
埼玉県	福岡県	の	市区	2
千葉県	佐賀県	町	区	3
東京都	長崎県	字	町	4
神奈川県	熊本県	丁	字	5
新潟県	大分県			6
富山県	宮崎県			7
石川県	鹿児島県			8
福井県	沖縄県			9
山梨県				
長野県				
岐阜県				
静岡県				

図 3.5: 内容判定用キーワード辞書

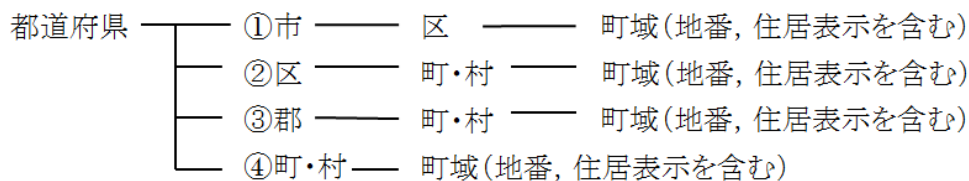


図 3.6: 住所の階層構造モデル

求人票の中には、1つの項目名に対して、対応する複数の内容が表記されているものがある。例えば、図 3.7(a) の求人票では、罫線の枠内に複数の内容が表記されている。これらの求人票では、罫線により内容間が区切られていないため、コンマによる文字列の分離ができない。そのため、原画像では別の行に記述された2つの文字列が結合する場合がある。

例えば図 3.7(b) では、内容「東京都港区芝浦1丁目2番1号シーバンスN館」と内容「山口県宇部市大字小串 1978-96」が結合していることがわかる。そのため、統一フォーマットへ変換するためには、これらの文字列を分離して整列する必要がある。本処理では、結合する内容間にはスペース、改行コードが存在する点や住所表記の先頭に郵便マーク（〒）が記入されている場合がある点に注目している。また、前述した内容候補の判定処理（方針2）を利用する。具体的には、以下の処理を行う。

1. 内容判定用キーワード辞書と文字列整理処理の適用結果を用いた、内容候補の判定処理（方針2）により、会社名や本社所在地の内容候補の記入位置を特定する。
2. 内容候補の記入位置に対して、「文字列整理処理の途中結果」中の相当する文字列を上書きする。
3. 文字コードに注目し、特定した内容候補の直後にあるスペース及び改行を判定し、その直前で文字列を分離する。「内容抽出数」に1を加算する
4. 〒マークの直前で文字列を分離する。
5. スペース、改行、ダブルコーテーションを除去する。

図 3.8 に、図 3.7(b) に示した文字列整理処理の適用結果に対して、内容候補間の分離処理を適用した結果を示す。図では、本社所在地についての内容候補を示す文字列が分離され、整列されていることがわかる。

東京本社 宇部本社	東京都港区芝浦1丁目2番1号 シーバンスN館 山口県宇部市大字小串1978-96
--------------	---

2行で表記された住所

(a) 求人票（入力画像）

2つの内容候補の文字列が結合
2005年4月採用東京本社宇部本社 東京都港区芝浦1丁目2番1号シーバンスN館山口県宇部市大字小串1978-96

(b) 文字列整理処理の適用結果

図 3.7: 別の行に記述された2つの内容が結合した例

2005年4月採用東京本社宇部本社 東京都港区芝浦1丁目2番1号シーバンスN館 山口県宇部市大字小串1978-96	分離
---	----

図 3.8: 内容候補間の分離処理の適用結果

また求人票の中には、内容のみが表記され、対応する項目名が表記されていない場合がある。例えば図3.9(a)の求人票では、会社名の項目名が表記されていない。これらの求人票に対して文字列整理処理を適用した場合、処理結果には項目名が格納されない（図3.9(b)）。そこで統一フォーマットへ変換するため、内容候補に対して適した項目名を付加する。現段階ではアルゴリズム簡易化の観点から、対応する項目名が表記されていない内容候補の記入位置を特定せず、前述した内容候補の判定処理（方針2）の適用結果に該当する全ての内容候補に対して項目名を付加している。

以下に、具体的な処理の流れを示す。

1. 内容判定用キーワード辞書と内容の分離処理の適用結果を用いた、内容候補の判定処理（方針2）により、会社名や本社所在地に関する内容候補の記入位置を特定する。
2. 内容候補の記入位置に対して直前の行に、その内容候補に適した項目名を付加し、内容候補と付加した項目名を抽出する。

図3.10に、図3.9(b)に示した内容候補間の分離処理の適用結果に対して、内容候補に対する項目名の付加処理を行った結果を示す。図では、内容に適した項目名「会社名」が付加されていることがわかる。

書 類 送 付 先 お 問 合 せ 先	〒460-0008 名古屋市中区栄2-6-1 白川ビル別館5階 株式会社 情報システム名古屋研究所 採用担当：技術総務グループ総務チーム
	TEL：052-201-0431 FAX：052-232-0028

(a) 求人票

書類送付先お問合せ先	内容のみ
〒460-0008名古屋市中区栄2-6-1白111ビル別館5階	

(b) 内容候補間の分離処理の適用結果

図 3.9: 対応する項目名が表記されていない例

書類送付先お問合せ先	付加した項目名
本社所在地	
〒460-0008名古屋市中区栄2-6-1白111ビル別館5階	

図 3.10: 内容候補に対する項目名の付加処理の適用結果

3.3 項目名候補と対応する内容の抽出

3.2 節で述べた「方針1」に定めた内容候補の判定を行う。

方針1では、項目名を表す文字列を特定する必要がある。そこで本研究では、求人票中に表記される項目名の表現・記述方法における規則性に注目し、求人票100件中に表記された項目名を登録した単語辞書（以下、項目名判定用キーワード辞書）を独自に作成した。図3.11に、単語辞書の中身を示す。この辞書と内容候補に対する項目名の付加処理の適用結果間で、文字列の一致判定を行うことにより、項目名候補を判定する。しかし、項目名以外の文字列の一部が項目名候補として判定される場合があるため、項目名独自の特徴に注目した。求人票100件を対象に項目名の特徴を調査した結果、「項目名は原則的に漢字表記の単語であり、場合によりフリガナや括弧が付加される」ことがわかった。

図3.12に求人票における項目名表記の例を示す。以上の点を考慮し、項目名判定用キーワード辞書と項目名の単語判定を利用した内容候補の判定方法（方針1）を以下に示し、図3.13で図示する。

- 1. 項目名判定用キーワード辞書と内容候補に対する項目名の付加処理の適用結果間で、文字列の一致判定を行う。
- 2. 一致した文字列を始点とし、右へ文字コードを探索し、スペースまたは改行コードが発見された地点を「文字列の終点」とする。
- 3. 文字列が、前述した項目名の特徴を示す場合、文字列を項目名候補として判定し、対応する内容と共に抽出する。

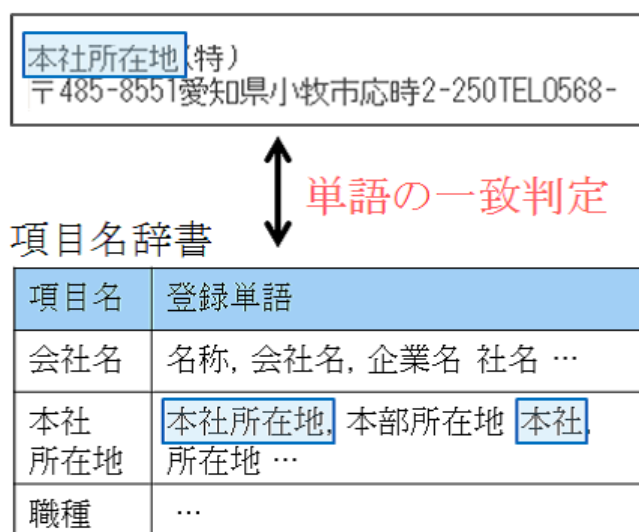
項目名の単語表記を考慮しない処理では、項目名判定用キーワード辞書の単語登録数の増加に伴い、項目名を内容と誤判定する数が増加する傾向があった。そこで、項目名独自の特征に注目し、項目名の単語表記を考慮することにより、内容との誤判定の数を低減できた。

<企業名>	<職種>	<待遇>	<仕事内容>
名称	職務内容	給与	事業内容
会社名	採用職種	初任給	仕事内容
企業名	募集職種	諸手当	業種
社名	求人職種	昇給・賞与	
	職種	昇給・賞与	
	仕事内容	賞与	<終了>
<所在地>		昇給	
本社所在地		福利厚生	
本部所在地	<採用>	勤務時間	
本社	選考方法	終業時間	
所在地	募集学科	休憩時間	
	採用学科	休日	
	応募資格	休暇	
	区分	休日休暇	
		有給休暇	

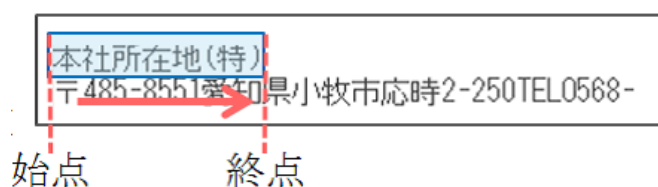
図 3.11: 項目名判定用キーワード辞書

フリガナ 企業名	フリガナ 社名	企業名 代表者	名 称	名 称 所在地 (本社)	ふりがな 名 称
フリガナ		本 社			
本社所在地	本 社	本 社 所 在 地	所在地		所 在 地

図 3.12: 求人票における項目名表記

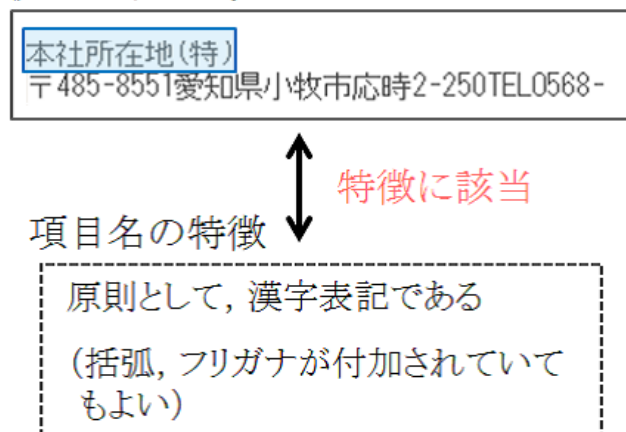


(a) 手順1



(b) 手順2

統一フォーマット



(c) 手順3

図 3.13: 内容の表記に注目した内容候補の判定手順

3.4 内容の選択

前述した内容候補の整列と抽出処理と項目名候補と対応する内容の抽出処理により、それぞれ抽出した項目名候補と内容候補から作成した「項目名候補と内容候補を抽出した文字列」を図 3.14 に示す。図から明らかなように、項目名候補と内容候補を抽出した文字列では、項目名候補から内容候補の抽出が容易である。項目名候補と内容候補を抽出した文字列では、求人票に表記された項目名と内容候補の整列と抽出処理により、作成された項目名が競合する場合がある（図 3.15）。そのため項目名を選択して、抽出する内容を決定する必要がある。具体的には、以下の処理を行う。

1. 項目名判定用キーワード辞書と項目名候補と内容候補を抽出した文字列による、内容候補の判定処理（方針 1）により、項目名候補を特定する。
2. 特定した項目名候補の中に、求人票に表記された項目名がある場合、その文字列の直後の文字列から「内容抽出数」個の文字列（内容）を抽出する。
3. 特定した項目名候補の中に、求人票に表記された項目名がない場合、3.2.2 節で付加した項目名の直後にある文字列から「内容抽出数」個の文字列（内容）を抽出する。
4. 複数の内容候補が抽出候補となった場合、最初に抽出候補と判定された内容候補（内容）のみを抽出する。

図 3.16 に、作成した統一フォーマットの例を示す。

会社名	項目名
株式会社	
資本金	内容
182億円	
本社所在地	
〒222-8558神奈川県横浜市港北区大豆戸町275番地	

図 3.14: 項目名候補と内容候補を抽出した文字列

会社名(特)	
株式会社	
本社	求人票に表記された項目名
所在地	
東京都港区西麻布2丁目26番地30号TEL03(3406)2111(代表)	
原貝Ⅱ的には、修士卒1978年4月1日以降出生者博士卒1975年	
不問の方	
東京本社	
もしくは工場(こちらで指定されています。)	
本社所在地(特)	付加した項目名
〒106-8620東京都港区西麻布2丁目26番地30号	
TEL : 03-3406-26g4FAX : 03-3406-22gg	

図 3.15: 項目名候補が競合した例

項目名	企業ID,企業名,所在地名,勤務地,初任給(大学),初任給(大学院) ↓
内容	1, 株式会社, 〒222-8558神奈川県横浜市港北区大豆戸町275番地

図 3.16: 統一フォーマット (CSV 形式)

第4章

評価実験

4.1 実験概要

提案法の有効性を確認するため、以下の観点に注目して、それぞれ評価実験を行った。

1. 提案法による適切な内容の抽出精度
2. 手入力による企業データベース作成方法と比較した提案法の運用効率

4.2 提案法による内容の抽出精度についての評価

求人票の文書画像を対象とし、提案法によりどの程度の精度で求人票中の内容を適切に抽出できるか確認するため、評価実験を行った。

4.2.1 評価用文書画像

2.2 節で述べた求人票 140 件（140 社）の文書画像を使用した。本学電気電子工学科に対して送付された 2006 年卒大学生向けの求人票の総数は 234 枚である。その内、本論文で研究対象とする「全体が罫線による表構造」の求人票は 140 枚であり、求人票全体の 59.8%を占める。本実験では、この 140 枚の求人票を文書画像化して使用した。

4.2.2 評価方法

特定の項目について、提案法により内容を適切に抽出できた求人票の件数を「抽出件数」、評価に使用する求人票の総数を「求人票数」とし、「抽出率」を以下の 4.1 式を用いて算出した。項目の中でも重要視される「会社名」、「本社所在地」に焦点を当て、それぞれ抽出率を算出した。

$$(\text{抽出率}) = \frac{(\text{抽出件数})}{(\text{求人票数})} \times 100[\%] \quad (4.1)$$

4.2.3 結果と考察

表 4.1 に評価実験の結果を示す．会社名は 99.3 % (139/140)，本社所在地は 90.7 % (127/140) の抽出率が得られた．図 4.1，4.2，4.3 に内容の抽出の成功例を示す．

図 4.1 では，求人票中の「項目名のない会社名」を適切に抽出できている．また図 4.2 では，求人票中に会社名のものと判断しにくい項目名「商号」が表記されているが，会社名を適切に抽出できている．さらに図 4.3 では，図 4.1，図 4.2 と異なり，会社名と本社所在地が左右の位置に表記されており，本社所在地が住所表記ではないが，会社名と本社所在地を適切に抽出できている．以上のように，文書構造（罫線の割合や文字列の表記位置）や表記された文字列（項目名や内容）が多様な場合においても，会社名や本社所在地が適切に抽出できていることがわかる．

表 4.1: 内容の抽出率に関する実験結果

	求人票数 [件]	抽出件数 [件]	抽出率 [%]
会社名	140	139	99.3
本社所在地	140	127	90.7

(株) [会社名] 募集要項		会社名
(2005 年 3 月卒業(修了) 予定者向け)		本社所在地
ふりがな 会 社 名	株式会社 [会社名]	(代表者) 取締役社長 [代表者]
分社・合併等 の設立内容	2002 年 12 月に IBM 社の Storage Technology Division が会社分割し、2003 年 4 月 1 日に [会社名] のハードディスクドライブ (HDD) 事業部門と統合し、設立。	
株式比率	[株式比率] 70%他	
本社所在地	〒256-8510 神奈川県小田原市国府津 2 8 8 0 番地 TEL0465-48-1111(大代)	

(a) 求人票

(株) [会社名] 募集要項
〒256-8510神奈川県小田原市国府津2880番地TEL0465-48-1111(大代)

(b) 抽出した内容

図 4.1: 内容の抽出の成功例 1

商 号	株式会社 [REDACTED]	会社名
所在地	三重県松阪市広陽町38番地 (松阪中核工業団地)	本社所在地

(a) 求人票

株式会社 [REDACTED] 所在地 三重県松阪市広陽町38番地 (松阪中核工業団地) 資本金

(b) 抽出した内容

図 4.2: 内容の抽出の成功例 2

		会社名
		本社所在地
事業所名	株式会社 [REDACTED]	本 社 2拠点 (東京・名古屋)
採用部署	リクルートセンター 中日本グループ	エンジニアリ ングセンター 名古屋、岡崎、他
所 在 地	〒451-0075 名古屋市西区康生通2-20-1	全国主要都市39拠点
書類提出先	同上	株 式 東証1部 名証1部

(a) 求人票

株式会社 [REDACTED] 2拠点(東京・名古屋)

(b) 抽出した内容

図 4.3: 内容の抽出の成功例 3

また、図 4.4, 4.5, 4.6 に内容の抽出の失敗例とその求人票及び文字認識結果を示す。

図 4.4(c) では、付加した項目名「本社所在地 (特)」により、内容「(空白)」が抽出されている。本来は、内容「株式会社 (旧社名)」を抽出することが望ましい。この誤抽出の原因として、図 4.4(b) に示した文字認識結果では、文字認識ソフトの誤認識により、文字認識結果上で求人票の表記構造 (図 4.4(a)) を再現できていない点が挙げられる。

図 4.4(a) に示した求人票では、項目名「社名」に対して内容「株式会社 (旧社名)」が記述されている。しかし文字認識結果では、項目名を「社名...代表者...」と誤認識したため、3.2.2 節 (方針 1) で述べた判定により、会社名に関する項目名として「社名...代表者...」を判定せず、付加した項目名を内容の抽出処理に利用している。

一方、文字認識結果では、文字列「(株)」と会社名を示す文字列の間にコンマが表記されている。そのため、これらの文字列は文字列整理処理により分離され、項目名の付加処理により内容「(株)」に対して項目名「会社名 (特)」が付加される。つまり文字認識結果上では、項目名とその内容が連続して表記されず、2.3 節で述べた「求人票における項目名と内容の位置関係」が利用できない。そのため、適切な内容を抽出できなかったと考えられる。

改善方法としては、求人票に特化した文字認識ソフトを開発し、表構造を正しく認識させることが考えられる。また使用する文字認識ソフトに、最も適する画像データの仕様を細かく調査することなどが挙げられる。

会	社名	株式会社 XXXXXXXXXX (旧社名 XXXXXXXXXX)	採用実績	専攻内訳 (単位:名)	2004年度 電気
	代表者	取締役社長 XXXXXXXXXX		募集人員	事務
	本社	大阪府高槻市古曽部町2-3-21 (最寄の駅)JR高槻駅、阪急高槻市駅より徒歩15分			

(a) 求人票

Text in blue box: 文字列が分離

Text in pink box: 項目名どうしが結合

Other visible text: 求人票<'2005年3月卒対象>, W柵, 株, 会社, 社名, 代表者, 株式会社

(b) 文字認識結果

内容の抽出不足

(c) 抽出した内容

図 4.4: 内容の抽出の失敗例（文字認識処理による影響）

また図 4.5(b) では、付加した項目名「本社所在地 (特)」により、内容「大阪本社: 〒590-8501 大阪府堺市鉄砲町 1 番地 事業支援センター人事グループ 担当 TEL:0120-27-3007() ホームページ:<http://www.daiceLcojp/Saiyo>」が抽出されている。本来は、内容「大阪本社 (堺市) ~ 西播磨研修センター (兵庫県赤穂郡)」を抽出することが望ましい。この誤抽出の原因として、主に 2 点が挙げられる。

1点目は、項目名「事業所 及び 所在地」が3.2.2節で述べた項目名の単語判定基準を満たしておらず、項目名として判定されなかったことである。そのため図4.5(a)に示した求人票では、本社所在地の項目名が判定されず、処理により付加した項目名を用いて内容の抽出処理が行われた。

2点目は、項目名「事業所 及び 所在地」に対応する内容「大阪本社(堺市)～西播磨

研修センター(兵庫県赤穂郡)」が住所表記ではないことである。本社所在地を対象とした項目名の付加処理では、住所表記に基づいて項目名を付加するため、求人票下部の内容「大阪本社: 〒590-8501 大阪府堺市鉄砲町1番地 事業支援センター人事グループ 担当 TEL:0120-27-3007(フリーダイヤル) ホームページ:http://www.daiceLco.jp/Saiyo」に項目名が付加され、内容が抽出された。

そこで今後は、項目名の特徴として挙げた「原則的に漢字表記である(括弧やフリガナが付加しても良い)」に加えて、新たな項目名の判定方法を検討する必要がある。

③事業場 及び 所在地	大阪本社(堺市)	東京本社(東京都)	名古屋支社(名古屋市)
	大阪営業所(大阪市)	福岡営業所(福岡市)	堺工場(堺市)
	神崎工場(尼崎市)	姫路製造所(姫路市)	播磨工場(兵庫県揖保郡)
	新井工場(新潟県新井市)	大竹工場(広島県大竹市)	
	総合研究所(姫路市)	筑波研究所(茨城県つくば市)	
	西播磨研修センター(兵庫県赤穂郡)		

連絡先

大阪本社: 〒590-8501 大阪府堺市鉄砲町1番地
事業支援センター人事グループ 担当 [REDACTED]
TEL: 0120-27-3007 (フリーダイヤル)
ホームページ: http://www.daiceLco.jp/saiyo

(a) 求人票

工業株式会社	内容の誤判定
大阪本社: 〒590-8501大阪府堺市鉄砲町1番地事業支援センター人事グループ担当 [REDACTED] TEL: 0120-27-3007(フリーダイヤル)ホーム	

(b) 抽出した内容

図 4.5: 内容の抽出の失敗例(項目名の誤判定)

図 4.6(b) では、項目名「東京本社」により、内容「〒105-8552 東京都港区海岸2丁目1番7号 日本板硝子東京ビル TEL(03)5443-9528」のみが抽出されている。本来は、項目名「大阪本社」に対する内容「〒541-8559 大阪市中央区北浜4丁目7番28号 住友ビル2号館 TEL(06)6222-7515」についても抽出することが望ましい。

この原因として、内容の抽出処理では、抽出に利用する項目名を1つに限定していた点があげられる。今後は、図 4.6(a) に示した求人票と同様の、複数の項目名が表記された求人票に対応するため、抽出する内容の数を増加させ、内容の抽出率を評価する必要がある。

東京本社	〒105-8552 東京都港区海岸2丁目1番7号 日本板硝子東京ビル TEL(03)5443-9528
大阪本社	〒541-8559 大阪市中央区北浜4丁目7番28号 住友ビル2号館 TEL(06)6222-7515

(a) 求人票

株式会社	内容の抽出不足
〒105-8552東京都港区海岸2丁目1番7号日本板硝子東京ビルTEL(03)5443-9528	

(b) 抽出した内容

図 4.6: 内容の抽出の失敗例（内容の抽出数の不足）

4.3 手入力の場合と比較した提案法の運用効率についての評価

手入力により企業データベースを作成した場合と比較し，提案法を利用することにより，どの程度の作業時間が削減できるか確認するため，評価実験を行った。

4.3.1 評価用文書画像と入力対象項目

2.2 節で述べた求人票 140 件（140 社）の文書画像を使用した。また，企業データベースへの入力対象とした項目は，現状で想定している就職相談システムでの検索項目に基づいて決定した。具体的には，会社名，本社所在地，職種，勤務地，福利厚生，資本金，初任給，従業員数を入力対象とした。

4.3.2 評価方法

前述した入力項目を企業データベースへ手入力するのに要する時間は，求人票の文書画像 50 件を使用し，10 件ごとに計 5 回計測した。なお，求人票中の内容を企業データベースへ手入力のに要する時間は個人差があるので，2 名の人員（各 20 件・30 件）による入力時間を計測し，10 件あたりの平均値を算出した。

また現状では，市販の文字認識ソフトにより作成した文字認識結果に対して，企業データベースへの自動登録手法を適用している。そこで，提案法による企業データベースの作成時間は，市販の文字認識ソフトにおける文字認識結果の作成時間と企業データベースへの自動登録処理時間の和により算出した。なお，市販の文字認識ソフトにおける文

字認識結果の作成時間と企業データベースへの自動登録処理時間の計測では，入力項目を会社名と本社所在地に限定し，手入力での実験で使⽤した求人票の文書画像 50 件を使⽤して 10 件ごとに計 5 回計測した．

4.3.3 結果と考察

表 4.2 に，求人票中の内容を企業データベースへ手入力するのに要する時間を示す．表より，手入力による企業データベースの作成には，10 件あたり 28 分 52 秒が必要となる．本学電気電子工学科における平成 24 年の就職活動での運用を想定すると，1 年間に送付された求人票は 501 枚であり，企業データベースの作成には 24 時間 3 分が必要となる．したがって，手入力による企業データベースの作成には膨大な時間が必要となり，就職相談システムの運用上，現実的な手段ではない．

表 4.2: 求人票中の内容を企業データベースへ手入力するのに要する時間
(求人票 10 枚あたり)

試行数 [回目]	入力時間
1	26 分 52 秒
2	29 分 44 秒
3	26 分 59 秒
4	32 分 56 秒
5	27 分 48 秒
平均	28 分 52 秒

表 4.3 に，市販の文字認識ソフトにおける文字認識結果の作成時間と企業データベースへの自動登録処理時間，それらの合計時間を示す．表より，提案法による企業データベースの作成には，10 件あたり約 7 分 2 秒が必要となる．本学電気電子工学科における平成 24 年の就職活動での運用を想定すると，求人票の文書画像 501 件あたり約 5 時間 52 分が必要となる．また，企業データベースへの自動登録処理時間は，就職相談システムの運用上，問題のない処理時間といえる．しかし，計測された時間は入力項目を会社名，本社所在地に限定したものであるため，就職相談システムの検索項目を全て入力する時間

を考慮する必要がある。

表 4.3: 提案法による企業データベースの作成時間（求人票 10 枚あたり）

試行数 [回目]	文字認識結果の 作成時間	企業データベースへの 自動登録処理時間	合計時間
1	7 分 26 秒	1.95 秒	7 分 28 秒
2	6 分 59 秒	1.94 秒	7 分 01 秒
3	6 分 39 秒	1.99 秒	6 分 41 秒
4	6 分 58 秒	1.94 秒	7 分 00 秒
5	7 分 00 秒	1.96 秒	7 分 02 秒
平均	7 分 00 秒	1.96 秒	7 分 02 秒

入力項目を会社名，本社所在地に限定した場合，処理全体で短縮される時間は「文字認識ソフトの修正機能を利用した，文字認識結果における誤読文字の修正時間」と「企業データベースへの自動登録処理時間」だと考えられる．そこで，文字認識ソフトの誤読文字を修正せずに文字認識結果を作成する時間を計測した．なお，使用する求人票の文書画像と文字認識結果の作成時間の計測方法は表 4.3 と同様とした．表 4.4 に，誤読文字を修正せずに文字認識結果を作成する時間を示す．

表 4.4: 誤読文字を修正せずに文字認識結果を作成する時間（求人票 10 枚あたり）

試行数 [回目]	作成時間
1	7 分 20 秒
2	6 分 05 秒
3	6 分 16 秒
4	5 分 51 秒
5	6 分 02 秒
平均	6 分 19 秒

表 4.3 と表 4.4 より、「市販の文字認識ソフトにおける文字認識結果の作成時間」と「文字認識ソフトの修正機能を利用した、文字認識結果における誤読文字の修正時間」の差を求めることにより、文字認識ソフトを利用した誤読文字の修正時間が算出できる。その結果、会社名と本社所在地の修正時間は、求人票の文書画像 10 件あたり約 41 秒程度であり、企業データベースへの入力項目が増加した場合でも、この数倍程度の時間であると推定できる。なお現状では、企業データベースへの入力項目が増加した場合の企業データベースへの自動登録処理時間が推定できないため、提案法による企業データベースの作成時間には考慮しない。

以上のことから、手入力による企業データベースの作成には、求人票の文書画像 501 件あたり約 24 時間 3 分が必要であるのに対し、提案法による企業データベースの作成には、約 5 時間 52 分程度が必要となる。したがって、提案法は手入力の場合と比較して、企業データベースの作成時間を 5 分の 1 程度に削減でき、提案手法は有用であるといえる。

今後は、求人票の他の項目についても企業データベースへの自動登録を行い、更に処理時間を短縮する手法についても検討する必要がある。例えば、文字認識ソフトの誤読文字に対する修正時間を短縮するため、企業データベースへの自動登録処理で使用するキーワード辞書に対して、文字認識誤りを考慮する拡張したキーワード（ビルの場合、ビルやビノレなど）を登録する方法など挙げられる。また、求人票に特化した文字認識ソフトを開発し、企業データベースへの自動登録処理を込みこむ（図 4.7）ことにより、処理時間の削減が期待できる。

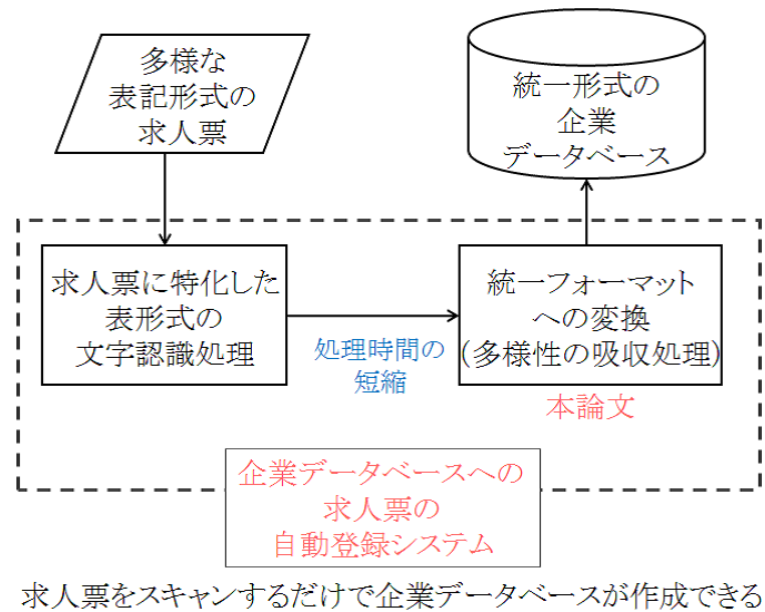


図 4.7: 求人票画像から企業データベースへの自動登録システムの将来像

4.4 本手法の問題点

本手法では、求人票の最も基本的な特徴である「表形式の基本構造を有し、記入された項目名と内容が左右で隣接し、対応関係を示す」点に基づいて処理している。本研究では、求人票中の会社名、本社所在地、勤務地、資本金、従業員数を対象とし、抽出を試みてきた。その後、より学生の要望を満たすために、職種、初任給についても企業データベース情報へ追加することとなった。しかし、求人票の「項目名と内容における左右の対応関係」の利用だけでは、職種、初任給を抽出できない場合がある。例えば図 4.8 では、項目名「初任給」に対応する内容が、複雑な表形式により表記されている。

そこで今後は、本手法に汎用性を持たせ、複雑な表構造や表記構造に対応するため、より詳細に求人票の表構造（文書構造）を解析（[20][21][22][23]）して利用する必要がある。

平成17年3月卒		求 人 票													
求 人 先	社 名	株式会社						私資 込本 済金	2億円		創 業	昭和44年6月	設 立	昭和62年7月	
	本 社 所在地	〒552-0001 大阪府大阪市福島区海老江1丁目1番31号 電話 06-6456-5200						年 商	107億円 平成15年3月実績		系 列	阪 神 電 気 鉄 道			
	事業本部	大阪市、鎌倉市						株 式	非上場		代 表 者		人 事 担 当		
	営業所	東京都、京都市、神戸市						従 業 員	480名		書 類 提 出 先	〒553-0001 大阪市福島区海老江1丁目1番31号 採用担当 電話 06-6456-5200			
	事業内容	1. コンピュータシステム(汎用、プロコン、EWS、クライアント・サーバ、パソコン)の販売、設計、開発、保守 2. マイコン応用機器の設計、開発、販売 3. インターネット関連事業 4. アウトソーシング事業													
採 用 条 件	採用 予 定 数	男 女	採用対象	大学院	大学	短大	高専	既卒		職 種 システムエンジニア セールスエンジニア プログラマ 電子機器設計技術者	初 任 給	男 女			
			理 系	可	15		5	不可				職種	技術	技術	技術
			前年実績	4	10	0	3		学歴			大学院卒	大学卒	高専卒	
	学部 採用学科 専攻	理系	理工系全学科					賞 与	初年度			次年度以降	昇 給	年1回 約2~4%	
	勤務時間	フレックスタイム制(コアタイム10:00~15:45)						年2回 約4ヶ月	年2回 約5ヶ月						
	休日	完全週休2日制(年間休日122日)、特別休暇						提出 書類	履歴書 成績証明書 卒業見込証明書 健康診断書						
	勤務予定地	大阪市、鎌倉市、東京都						携 帯 品	印鑑 筆記用具						
採 用 試 験	方 法	筆記(専門、一般常識) 小論文、適性試験、 健康診断、面接		会社説明会				他 の 条 件	平成17年3月 卒業見込みの者						
	締切日	随 時													
	試験日	随 時													
	場 所	本 社													

本社 電話 06-6456-5200
〒553-0001 大阪市福島区海老江1丁目1番31号
交通費は当社にて負担させていただきますので、事前に電話で連絡の上、お越し下さい。

図 4.8: 複雑な表構造を内包する求人票

第5章

おわりに

5.1 本論文のまとめ

本研究では，学生の就職活動の質を向上させ，中小企業の採用活動を改善する就職支援システムを提案し，システムへのデータ入力の手間を軽減する手法として，求人票中の文字列の自動データベース化手法を検討した．本論文では，求人票中の文字列を項目名と内容に分類し，ほぼ全ての求人票が内包する共通的な特徴として「表形式の基本構造を有し，記入された項目名と内容が左右で隣接し，対応関係を示す」点を基軸として，内容の抽出を試みた．また，一部の求人票の有する特徴として「1つの項目名に対して複数の内容が表記されている」点や「項目名がなく，内容のみが表記されている」点に注目して，適切な内容の抽出を試みた．

本手法を140件(140社)の求人票に適用した結果，会社名は99.3%(139/140)，本社所在地は90.7%(127/140)の抽出率が得られた．また，本手法による企業データベースの作成には，求人票501件あたり約5時間52分程度が必要となり，手入力の場合と比較して企業データベースの作成時間を5分の1程度に削減できる．

5.2 今後の課題

今後は本手法に汎用性を持たせ，より複雑な表構造や表記構造に対応するため，求人票の表構造(文書構造)をより詳細に解析して利用する必要がある．また，複数の項目名が表記された求人票に対応するため，抽出する項目数を増加させ，内容の抽出率の変化を評価する必要がある．さらに，その他の内容の抽出処理における失敗例の改善を行い，会社名や本社所在地以外の項目における内容の抽出方法を検討する必要がある．

将来的な課題として，提案システムのデータベースを蓄積し，システムの試験運用を行うことや求人票に特化した文字認識ソフトの開発などが挙げられる．

謝辞

本論文は、筆者が三重大学大学院工学研究科博士前期課程に在学中に行った研究をまとめたものである。本研究の遂行及び修士論文の作成にあたり、懇切丁寧な御指導と御督励を賜った本学地域イノベーション学研究科の鶴岡信治教授、本学工学研究科電気電子工学専攻の高瀬治彦准教授、川中普晴助教に深く感謝致します。また、本研究の研究方針の検討に協力していただいた三重大学名誉教授の三宅康二教授、東京理科大学工学部第一部電気工学科のブレイマチャンドラチンタカ助教に深く感謝致します。そして、貴重な時間をさいて本論文を査読していただいた本学工学研究科電気電子工学専攻の北英彦准教授に深く感謝致します。

最後に、日頃熱心に討論していただいた情報処理研究室の皆様と本論文をまとめるにあたり、ご助言、ご討論、その他お世話になりました全ての方々に感謝致します。

参考文献

- [1] 株式会社リクルート, リクナビ 2014:<http://job.rikunabi.com/2014/>, 2013
- [2] 株式会社マイナビ, マイナビ 2014:<http://job.mynavi.jp/2014/>, 2013
- [3] 株式会社学情, 学情ナビ 2014:<https://www.gakujo.ne.jp/2014/>, 2013
- [4] 株式会社日経 HR, 日経就職 Navi2014:<https://job.nikkei.co.jp/2014/top/>, 2013
- [5] 株式会社リクルート ワークス研究所, “ 第 29 回 ワークス大卒求人倍率調査 (2013 年卒) ”, 株式会社リクルートキャリア ホームページ: <http://www.sagar.jp/>, 2012
- [6] 四日市商工会議所 学生就職 PR センター, 三重就職 NAVI:<http://www.mie-snavi.net/action/main/index>, 2013
- [7] 株式会社マイナビ, “ 2013 年卒 マイナビ大学生就職意識調査 ”, マイナビ採用サポネット: http://saponet.mynavi.jp/enq_gakusei/ishiki/index.html, 2012
- [8] D. Okada:Development of “ Vocational Guidance Sytem ” Using the Job Database of Specific Department, 三重大学工学研究科電気電子工学専攻 平成 24 年度情報処理研究室中間発表予稿集, p.4, 2013
- [9] 長井達一郎, 塚本直子, 荒牧重登, 鶴岡智昭, “ 求人情報案内作成システムの開発 ”, 福岡大学工学集報, vol.80, pp.41-47, 2008
- [10] 高坂宜宏, “ 求人票閲覧システムの開発 ”, 釧路工業高等専門学校紀要, vol.44, pp.19-24, 2010
- [11] 朝倉利紀, 松崎大祐, 井上孝太郎, 徐海燕, “ 就職活動情報登録閲覧 Web システムの構築と運用 ”, 火の国情報シンポジウム 2009, vol.4, p.3, 2009
- [12] 三井所健太郎, 藤村直美, “ WEB インターフェースによる就職活動支援システムに関する研究 ”, 情報処理学会研究報告 (マルチメディア通信と分散処理研究会報告), vol.17, pp.1-6, 2009
- [13] 株式会社 PFU, 名刺ファイリング OCR:<http://www.pfu.fujitsu.com/meishi/>, 2013

- [14] メディアドライブ株式会社, 名刺認識ライブラリ v6.0: <http://mediadrive.jp/products/library/meishi/index.html>, 2013
- [15] 斎鹿尚史, 中村安久, 北村義弘, 森田敏昭, “名刺読み取りシステム”, 電子情報通信学会技術研究報告, vol.93, pp.41-48, 1993
- [16] 石谷康人, 中村敏弘, “OCR 誤りに対してロバストな文書画像を対象としたモデルベースと情報抽出”, 電子情報通信学会技術研究報告, vol.101, pp.123-130, 2002
- [17] 厚生労働省, 職業安定法施行規則, 厚生労働省令第一一四号: <http://law.e-gov.go.jp/htmldata/S22/S22F04101000012.html>, 2013(現在)
- [18] 富士通ミドルウェア株式会社, “表 OCR/文書 OCR for Excel & Word v.5.0 仕様書”, 2003
- [19] 財団法人 地方自治情報センター, “全国地方公共団体コード仕様”, LASDEC: <https://www.lasdec.or.jp/cms/1,7505,14.html>, 2008
- [20] 田端康人, 鶴岡信治, 木村文隆, 三宅康二, “表の構造理解のための罫線抽出と領域分け”, 電子情報通信学会技術報告, vol.90, pp.33-37, 1990
- [21] 浅野三恵子, 下辻成佳, “セル構造を用いた帳票識別”, 電子情報通信学会技術報告, vol.80, pp.131-138, 1997
- [22] 川崎洋治, 野村直之, 中川尚, “文書構造情報の抽出とメタデータ化”, 情報処理学会研究報告, vol.37, pp.43-50, 2003
- [23] 田仲正弘, 石田亨, “表構造の一般化に基づくオントロジの獲得”, 情報処理学会論文誌, vol.47, pp.1530-1537, 2006

発表論文リスト

国際会議

- (1) M. Shigenaga , S. Tsuruoka , H. Kawanaka , H. Takase :For “Vocational Guidance System” Automatic Database Assignment of Character Strings in Job Offer Form , Proc. of the Fourth Intl. Workshop on Regional Innovation Studies IWRIS2012 , pp.71-74 , 2012
- (2) M. Shigenaga , S. Tsuruoka , H. Kawanaka , H. Takase :From Job Offer Form Automatic Database Assignment of Character Strings to “Vocational Guidance System” , Proc. of The 2nd Intl. Symposium for Sustainability by Engineering at MIU IS-SEMU2012 , pp.245-248 , 2012

国内会議

- (1) 重永宜也 , 鶴岡信治 , 川中普晴 , 高瀬治彦 , “ 就職相談システムのための求人票中の文字列の利用法 ” , 平成 23 年度電気関係学会東海支部連合大会講演論文集 , G2-3 , 2011
- (2) 重永宜也 , 鶴岡信治 , 高瀬治彦 , 川中普晴 , “ 就職相談システム構築のための求人票中の文字列のデータベース化 ” , 平成 23 年三重地区計測制御研究会講演会講演論文集 , IP-04 , 2011
- (3) 重永宜也 , 鶴岡信治 , 川中普晴 , 高瀬治彦 , “ 就職相談システムのための求人票中の文字列のデータベース化 ” , 平成 24 年度電気関係学会東海支部連合大会講演論文集 予稿集 , B5-4 , 2012