データマイニングによる求人情報の分析

～企業と学生間のミスマッチ低減に向けて～

東京工芸大学工学部
コンピュータ応用学科
石嶋　秀太・宇田川　佳久

はじめに

就職は個人にとっても、企業および社会にとって重要な関心事である
最近の文部科学省の調査によると2013年3月に卒業した大学卒業者の求人倍率は1.27倍であるのに卒業生の2割以上が安定雇用に就いていない
→企業が採用基準に達する学生に出会えず採用予定数の確保に難航していると分析されている

はじめに②

大卒者は有名企業への就職希望の傾向があり
従業員1000人以上の企業の求人倍率0.73倍
従業員1000人未満の企業の求人倍率1.79倍
→学生と企業間でミスマッチが発生している

- ミスマッチ発生の原因：　企業情報は沢山あるが、自分の立場での把握が不十分
- 解決のために、データマイニングを使って、大量のデータから業界別の傾向を把握する研究を行っている

本研究の目的

① 商用のテキストマイニングツールの分析機能を確認すること

② 対象としたデータは，本学の企業説明会に参加した企業のプロフィール資料から取得した

③ 最終的には，求人データをインターネットからダウンロードし，データマイニングによる分析を実行し，結果をユーザに提示するシステムを目指している．

企業データの収集と加工

- 分析対象としたデータは，本学の企業説明会に参加した企業のプロフィール資料から、事業内容の紹介と学生へのメッセージ文を編集したものである
- 解析対象とした事業内容とメッセージは、通常使われている日本語の文章である
  ⇒　次のような編集、修正を行った

企業データの収集と加工②

編集、修正の例

1.企業固有の製品名、地名、部署名、親会社名、子会社名などは、データ作成時に削除した

2.標語のように、名詞形で終了している場合は、文章になるよう動詞を補った

3.文章の最初で、発話の主体（主語）が不明確な場合、「当社は」などを適宜補った

分析対象データの基本情報

	全体	情報	建築	電気	機械
企業数	79	37	13	12	17
総行数	158	74	26	24	34
平均行長(文字数)	68.3	73.5	58.6	64.6	67
総文数	396	197	62	58	79
平均文長(文字数)	27.3	27.6	24.6	26.7	28.8
延べ単語数	3101	1478	478	470	675

形態素解析機能

- 計算機で自然言語の文章を処理するためには、形態素（言語として意味を持つ最小単位）に分割し、それぞれの品詞を判別する　形態素解析を行った

形態素解析結果

サンプル文

基幹系　通信システムを中心としたソフトウェアの受託開発製品販売・ソリューション自社製品開発・研究

単語ID	見出し語	原形	置換語	品詞	係り先
286	基幹系	基幹系	基幹系	名詞	287
287	通信システムを	通信システム	通信システム	名詞	288
288	中心とした	中心	中心	名詞	289
289	ソフトウェアの	ソフトウェア	ソフトウェア	名詞	290
290	受託開発製品販売・	受託開発製品販売	受託開発製品販売	名詞	-1
291	ソリューション自社製品開発・	ソリューション自社製品	ソリューション自社製品	名詞	-1
292	研究	研究	研究	名詞	-1

形態素解析結果

- 名詞の発生数が事業内容、メッセージ共にもっとも多い
  →内容を説明する文が多いためだと考えられる
- 動詞の発生数が事業内容よりメッセージのほうが多い
  →感情の動きを表すものが多いからと考えられる

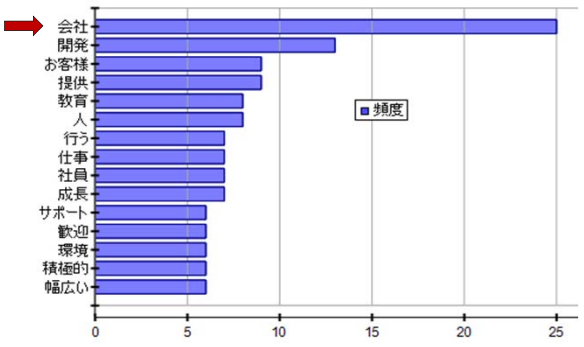
	事業内容		メッセージ	
	発生数	発生数の比率	発生数	発生数の比率
名詞	687	94.63%	1900	80.00%
動詞	16	2.20%	276	11.62%
形容詞	1	0.14%	74	3.12%
副詞	0	0.00%	61	2.57%
連体詞	2	0.28%	31	1.31%
接続詞	19	2.62%	29	1.22%
記号	1	0.14%	3	0.13%
間投詞	0	0.00%	1	0.04%
合計	726	100.00%	2375	100.00%

類義語辞書機能

- 会社、当社、弊社という用語が現れるが、これらは同じ意味で使われていると解釈できる。今回の分析で作成した辞書を適用することで　代表語に統合する

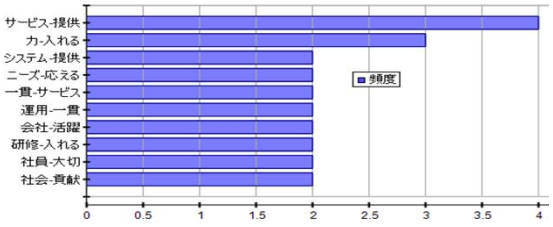
代表語	品詞	類義語1	類義語2	類義語3
パソコン	名詞	PC	パーソナルコンピューター	コンピュータ
会社	名詞	当社	弊社	
成長	名詞	スキルアップ	キャリアアップ	向上
サポート	名詞	支援	応援	支える
ものづくり	名詞	モノづくり		
先輩社員	名詞	先輩	先輩たち	
Android	名詞	android	アンドロイド	
教育	名詞	育てる	育成	

類義語辞書の適用後の単語発生頻度



係り受け頻度解析

- 係り受けに注目することで文章中に発生している単語同士の共起表現を推測することが可能となる
- 対象とした単語の品詞は名詞、動詞、形容詞、副詞、連体詞、接続詞である。

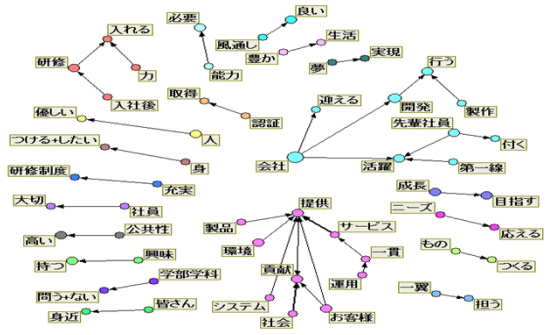


信頼度

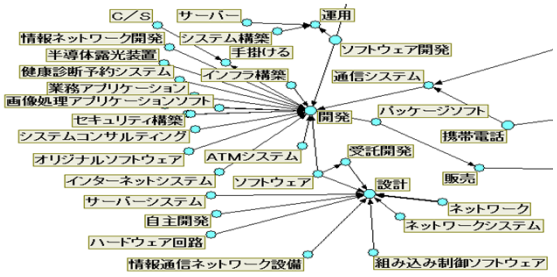
- 係り受け表現の発生頻度を、構成する単語を含む文章の数の相対値を基準として評価する
- n(x): 単語 x が発生した文章の数
- m(x,y): 単語 x と単語 y が同時に発生した文章の数
  - すなわち，m(x,y) は，文章全体で，単語 x と単語 y を含む係り受け表現の数を表す．
- 信頼度を次の式で定義する

  信頼度: Pxy= m(x,y) / n(x)

係り受け関係のグラフ表示の例



情報業界・事業内容の分析



情報業界・事業内容の分析

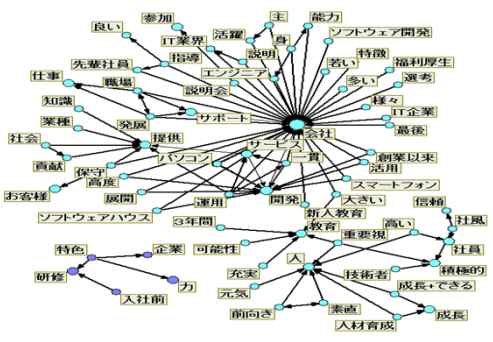
- 単語の発生頻度は，「開発」，「ソフトウェア開発」，「設計」，「携帯電話」という単語が多い

→ソフトウェア関連企業の求人情報であることから想定の範囲内の事象

「開発」から共起されている単語には「インフラ構築」があり，「設計」から共起されている語彙には「ネットワーク」，「情報通信ネットワーク設備」，「サーバシステム」がある

→企業としては，ソフトウェアやシステムを開発するために必要となる，インフラ関連技術にも注目していることが推測できる

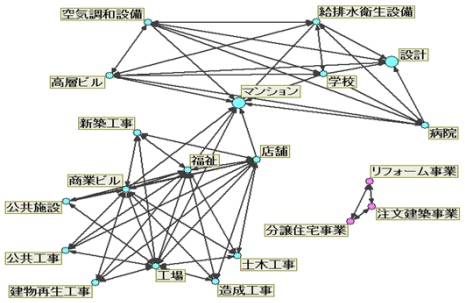
情報業界・メッセージの分析



情報業界・メッセージの分析

- 単語の発生頻度としては，「会社」，「開発」，「提供」，「サービス」，「お客様」，「教育」，「人」などが多い
- 上側は仕事に関連する事項であり，発生している単語の数と係り受けの数から，話題の主流を占めていることが分かる
- 下側は教育・研修に関する内容で，「入社前→研修」，「3年間→教育」，「元気，前向き，素直→人」などが読み取れる

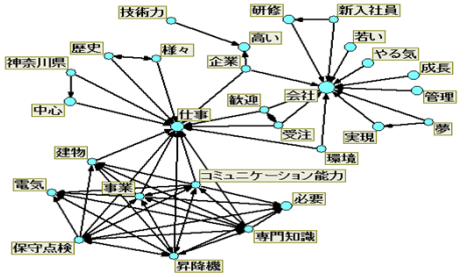
建築業界・事業内容の分析



建築業界・事業内容の分析

- 上側は，「マンション」，「学校」，「病院」の高層化とそれに関わる「空気調和設備」，「給排水衛生設備」が事業の対象になっていることが推測できる
- 下側は，工事の分類について言及している記述の単語の共起関係を表している
- 右下は，住宅事業に関するもので，注文建築に加えリフォーム事業を手掛けていることを示唆している

建築業界・メッセージの分析



建築業界・メッセージの分析

- 単語の発生頻度としては，「会社」，「仕事」が多い
- これに続いて，「高い」，「研修」，「やる気」，「成長」，「管理」，「必要」，といた単語が多いことが分かる
- 下側は仕事の内容に関する記述がみられる．それぞれの単語が相互に共起しているが，「専門知識」と並んで「コミュニケーション能力」が話題になっている

まとめ

- 求人情報を商用のテキストマイニングツールを使って分析した結果について述べた．
- 実験では，一部の文章に対し主語や動詞を補う修正をしたが，ほぼ原文のままの文章を解析した．
- 今回の分析により，単語や係り受け表現の発生頻度について統計的な観点で論じるためには，数百件のデータが必要であるとの知見を得た．
- 今後は，さらに多くの求人情報を使った分析を行い，企業と学生間のミスマッチ低減に向けた提言の抽出を試みる予定である

ご清聴ありがとうございました