# 部署役職テキストの自動分割

髙橋 寛治　奥田 裕樹
Sansan 株式会社 DSOC
{ka.takahashi, okuda}@sansan.com

## 1　はじめに

名刺は，個人における時間的な一点を切り取った断面図である．会社や個人を取り巻く様々な要因によって，名刺に書かれている情報は変化する．会社の合併や分割などによって会社名が変わることがあれば，組織再編や昇進などによって部署や役職が変わることもある．当然ながら転職や転籍によってその所属が変わることも考えられる．名刺交換は一度行った相手とは再度行わないのが通例であり，こういった名刺に記載される情報の変化の多くは補足し続けることが難しい．

名刺管理サービスを提供する Sansan では，部署や役職の情報を元に名刺交換相手の栄転など人事異動情報を配信する機能を，利用ユーザに対し提供している．更新された情報を取得する手段として，サービスに取り込まれた名刺のデータを用いる方法と，新聞社やウェブメディア等から発信される人事異動情報を用いる方法がある．名刺データにおいては部署や役職が分割され正確に記載されている一方で，新聞社やウェブメディアが公開している人事異動ニュースのテキストデータは人間の可読性を重視していることから，それらが区別されず連続した文字列として記載されていることが多い．

部署や役職は，企業が自由に決めることができる名前であることから，自由度が非常に高く多種多様な表現が存在する．部署や役職の変化を検知する方法において，職種などの分類から意味的な違いを大まかに捉えることは可能ではあるが，必ずしもすべての企業のルールを網羅した分類方法を作ることは困難であり，実際には部署や役職の文字列的な変化を捉え比較するしか方法はない．比較に際してはノイズの除去などの前処理に加えて，部署および役職が正しく構造化された状態で取得できていることが前提になる．そのため，構造化されていない部署および役職の情報を処理するうえで人間が手動で作業するには莫大な時間や労力がかかることから，任意のテキストから目的となる項目を抽出する情報抽出の技術が求められる．

本稿では，固有表現抽出の手法を応用し，系列ラベリングの問題として部署と役職を分割する方法を提案する．部署と役職が混在したテキストから名刺データ化により構造化されたデータを用いることでモデルを学習し，ウェブメディアが提供する部署役職が連続した文字列に対して分割を行う．

## 2　対象とする部署役職テキスト

部署と役職が区別されずに連続した文字列として記載されているテキストのことを以降では，**部署役職テキスト**と呼ぶ．今回のタスクで対象とする部署役職テキストは二つある．一つは，学習データとして用いる名刺に記載されている部署役職テキストから構造化された部署と役職である (以降では，**学習データ**と呼ぶ)．もう一つは，ウェブメディアが提供する人事異動ニュースの部署役職分割テキストである．本稿での分割対象である．

本章では，対象となる部署役職テキストおよび学習データについて説明する．

### 2.1　部署役職テキスト

今回対象とする部署役職テキストは，ダイヤモンド社が提供する人事異動ニュースである[1]．会社名や氏名とともに肩書きが記述されている．この肩書きに部署と役職が区別されずに連続した文字列，すなわち部署役職テキストが記述されている．

次に，部署役職テキストを模した作例を示す[2]．

- IT ソリューション事業部部長
- 取締役兼執行役員

[1] https://www.diamond.co.jp/go/jinji/
[2] 珍しい部署や役職だと特定できる可能性があるため，作例を記載する．

- 取締役／執行役員／常務
- 執行役員／@HogeLtd.
- 参与 (営業部・大規模担当)

このように，兼任が「兼」「／」「・」など様々な表記で記載されている．また「@」で企業名が記載されたり，参与の詳細情報は括弧で記述されたりと自由な記述である．まとめると，部署と役職以外の文字列も混入しているのが特徴である．

### 2.2 学習データ

名刺由来の部署役職テキストは人手により部署と役職に分割されている．表 1 に示すように，部署と役職は名刺作成者の考えた基準でチャンクされており，単語間には空白が挿入されていることが多い．このチャンクについて，詳細な部署検索機能を提供することを考えると，「徳島本部」と「営業部」が分割されていることが望ましい．

複数部署役職に所属している場合は，最初に出現する部署を部署として認定し，その他のテキストを役職とする．また部署か役職か判断がつかないテキストは役職としている．この判断基準は提供するサービスに依存している．

## 3 部署役職の分割

### 3.1 分割手法

CRF(Conditional Random Fields)[3] による系列ラベリングを行う．

部署と役職の系列の表現について説明する．IOB2 タグ [6] を用いることで，単語ごとに部署と役職ラベルを付与する．一つの部署を構成する要素として，分割可能な部署が連なる場合がある．例えば，「徳島本部 営業部」のように「徳島本部」と「営業部」の二つのチャンクとする．IOB2 タグにならってチャンクの先頭は B(Beginning) タグ，B タグに後続するトークンを I(Inside) タグとする．タグに部署と役職の情報を付与するため，「B-部署」というようなタグとする．

今回は O(Outside) タグを利用しない．タグは B・I と部署・役職の組み合わせで全部で 4 種類となる．

単語ごとにラベル付けを行うこと，および素性抽出のために，形態素解析を適用する．「徳島本部営業部」

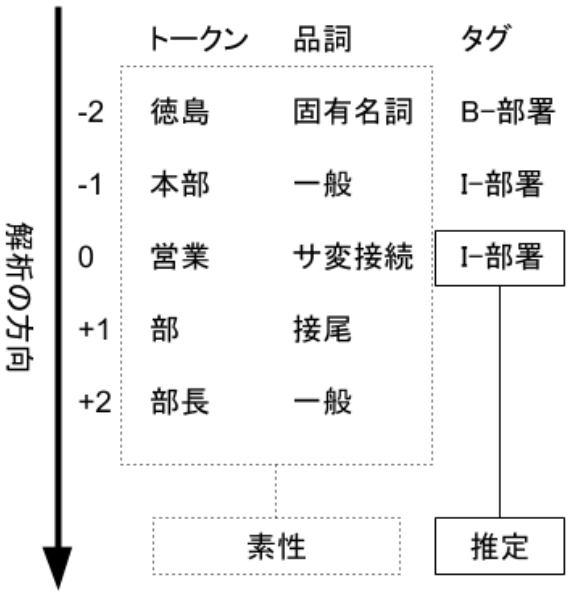


図 1: 素性とタグ

を「徳島 本部 営業 部」と分割したものに対して，B タグと I タグを付与する．タグを付与した例を図 1 に示す．

次に，素性について説明する．部署役職テキストの文頭には BOS(Beggining Of Sentence)，文末には EOS(End Of Sentence) を挿入する．BOS，EOS を含めた系列に対して素性抽出を行う．素性は，対象としている単語に対して，その単語も含めた前後 2 単語の表層形および品詞の細分類 1(IPADic での固有名詞や接尾などの属性) を用いる (図 1)．また，素性にバイアス項を挿入する．ここで，品詞を用いなかったのは，ほとんどが名詞だからである．事前に予測したタグは利用しない．すなわち，それぞれのトークンに対しての分類問題を解いている．

名刺由来の部署役職テキストと人事異動ニュースの部署役職テキストでは，兼任の表記が異なる．学習データの加工して差異を学習させるのではなく，スペースを「兼」と置換する前処理として対応する．

### 3.2 実利用のための分割精度の評価方法

適合率や再現率で分類器の性能を評価することは一般的である．提供サービスでは部署役職テキスト内の一件目の部署と役職を採用している．例えば，「執行役員 兼 人事部長」は「執行役員 兼 人事部長」が一つの役職テキストとしている．そのため，実際のデー

表 1: 名刺由来の部署役職の例

| 部署役職テキスト | 部署 | 役職 |
|---|---|---|
| 徳島本部 営業部 部長 | 徳島本部 営業部 | 部長 |
| 代表取締役 | | 代表取締役 |
| マーケティング部 | マーケティング部 | |
| 事業企画部 兼 マーケティング部 次長 | 事業企画部 兼 マーケティング部 | 次長 |
| 徳島支社 編集長 編集記者 | 徳島支社 | 編集長 編集記者 |
| 徳島資材課長 兼 香川資材課長 | | 徳島資材課長 香川資材課長 |

タに対する評価では，実サービスで求められる結果と一致しているかどうかを評価する．

## 4 実験

学習データを用いて部署役職の系列をラベリングするモデルを学習させる．学習データに対する性能評価と部署役職テキストに対する性能評価を行う．学習データに対する性能評価は適合率，再現率，F 値と精度で行う．部署役職テキストに対しては実利用を考えた分割精度を評価する．部署役職分割モデルは，学習データ 3000 件を対象に 5 分割交差検証で評価する．

CRF による系列ラベリングの学習には，crfsuite[5] を用いる．パラメータはデフォルト値を用いる．形態素解析には MeCab(IPADic)[3] を用いる．

## 5 結果と考察

結果や考察で示す例は，実際の部署役職テキストに似せて筆者が作った，架空の部署役職テキストである．

### 5.1 学習データ (同じドメイン)

学習データに対しての 5 分割交差検証した結果のそれぞれのタグの適合率，再現率，F 値および件数の平均を表 2 に示す．

それぞれの評価について平均を取ると，適合率は 0.90，再現率は 0.91，F 値は 0.91 となった．概ね 9 割程度の性能である．実利用としては，すべてのタグが正解かどうかが重要であり，その精度は 0.78 であった．

[3] http://taku910.github.io/mecab/

表 2: 5 分割交差検証の各タグの平均

| タグ | 適合率 | 再現率 | F 値 | 件数 |
|---|---|---|---|---|
| B-部署 | 0.94 | 0.95 | 0.94 | 879 |
| I-部署 | 0.97 | 0.95 | 0.96 | 583 |
| B-役職 | 0.85 | 0.89 | 0.87 | 1866 |
| I-役職 | 0.85 | 0.86 | 0.86 | 658 |

誤り事例をできるかぎり模倣した作例を示す．スペースはトークン区切り，カンマはチャンクの区切りを示す．

**入力** 徳島 支社 企画 営業 部 主任

**推定** B-部署 I-部署 B-部署 I-部署 I-部署 B-役職

**正解** B-部署 I-部署 I-部署 I-部署 I-部署 B-役職

「企画営業部」という部署に関して，「営業企画」という部署が頻出することから誤ったと考えられる．低コストでの解消方法としては「営業企画」という単語を辞書に追加することが考えられる．

### 5.2 部署役職テキスト (異なるドメイン)

部署役職テキスト 100 件に対しての 92 件正解であった．学習データに対しての精度が 0.78 であったため，正解が多い理由について考察する．

名刺由来のデータは多種多様な部署や役職が含まれている．一方で，今回対象とした部署役職テキストは，会社の役員や部門長，部長が対象となっている．そのため，「執行役員/マーケティング事業部本部長」といった部署役職テキストの場合，「マーケティング事業部」という部署が含まれているが，「執行役員」が一つ目の役職となり，これに対応する部署がないため，与え

られたすべてのテキストが役職となる．実利用を考慮した評価ではスコアが高くなる．

よく見られた誤り事例は，「徳島総支社長」といった場合に部署「徳島」と役職「総支社長」に分かれてしまう．これは「徳島総支社長」という役職である．誤り事例を修正したものを学習データとして追加することで，改善できるのではないかと考えている．

## 6 関連研究

与えられた単語列に対して，順番にラベルを予測する系列ラベリング [6] は，固有表現抽出や形態素解析などをタグ付けと見なして幅広く利用されている．例えば，形態素解析では文字やトークン単位でラベル付けの問題として解くことで適用されている [4, 8]．チャンキングにも応用できるためフレーズ抽出としても利用される [2]．このようにチャンク同定と見なして，様々な問題に対して応用されている．

人物に関連する情報抽出の取り組みでは，履歴書からの系列ラベリングによる情報抽出 [1] や Wikipedia の人物情報ページから属性を抽出する試み [7] が行われている．これらは，部署役職テキストのように出現する系列が限定的ではなく，テキスト中からの固有表現の抽出である．

本稿では，これまでに見られなかった部署役職テキストに対して，系列ラベリング問題として部署役職に分割した．

## 7 おわりに

本稿では部署役職テキストに対して，系列ラベリングの手法により部署と役職に分割する方法を示した．名刺由来の部署と役職のテキストを用いて部署役職分割器を学習し，部署役職テキストに対する分割を行った．結果，同じドメインに対しては精度が 0.78 だったが，部署役職テキストに対しては 0.92 だった．部署役職テキストに対しては，ドメインの異なりが要因で分類が簡単な部署役職テキストが多かったためである．今後は，見つけた誤りを効率よく改善する方法が必要となる．

## 参考文献


[1] Akihiro Katsuta, Hutama Adhi Hanjaya, Sommath Asati, Sorami Hisamoto, Kazuma Takaoka, and Yoshitaka Uchida. Infomation extraction from english & japanese rsum with neural sequence labelling methods. In *Proceedings of the Twenty-fourth Annual Meeting of the Association for Natural Language Processing*, pp. 1007–1010, 2018.

[2] 工藤拓, 松本裕治. Support vector machine を用いた chunk 同定. 自然言語処理, Vol. 9, No. 5, pp. 3–21, 2002.

[3] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[4] Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Nguyen, and Quang-Thuy Ha. Vietnamese word segmentation with crfs and svms: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pp. 215–222. Tsinghua University Press, 2006.

[5] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[6] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pp. 173–179, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[7] 関根聡, 小林暁雄, 安藤まや, 馬場雪乃, 乾健太郎. Wikipedia 構造化データ「森羅」構築に向けて. 言語処理学会第 24 回年次大会発表論文集, pp. 765–768, 2018.

[8] 吉田辰巳, 大竹清敬, 山本和英. サポートベクトルマシンを用いた中国語解析実験. 自然言語処理, Vol. 10, No. 1, pp. 109–131, 2003.