企業名からの職種の推定

渕 武志

NTT情報通信研究所

fuchi@isl.ntt.co.jp

概要

企業名に含まれる文字列が表す職種の確率を用いて、職種を推定するシステムを試作した。企業名からの文字列の抽出方法として字種、語長、形態素などいくつかの方法を比較した結果、全ての形態素を無差別に用いる方法が最も高い精度を示した。

1. まえがき

漢字は表意文字であるため、固有名詞においてもある程度その漢字の意味が反映されていると考えられる。特に企業名の場合、その職種を表す言葉を含む場合が多いので、企業名だけから職種を推定することがある程度可能であり、実際にどの程度の精度で職種を推定できるのか興味のあるところである。そこで、我々は法人顧客のデータベースから、企業名を構成する文字列が表す職種の確率を算出し、これを用いて新規の企業名に対してどの程度職種の推定が可能であるかを実験した。本稿ではこの実験について報告する。

2. 実験方法

試作したシステムは、企業名を表す文字列を入力として、推定した職種を出力するシステムである。システムで用いる職種の体系は、職業別電話帳に用いられている職種の中分類であり、136の職種が含まれている。

2.1. 企業名の分解

システム内では、企業名から文字列が抽出されて用いられる。実験では、幾つかの抽出方法を使った場合の職種推定精度を比較する。以下にその方法を示す。

2.1.1. 文字による分解

単純に一文字毎に分解する。この場合、分解された文字を無差別に用いる方法と、漢字のみを用いる方法を試した。

2.1.2. 字種による分解

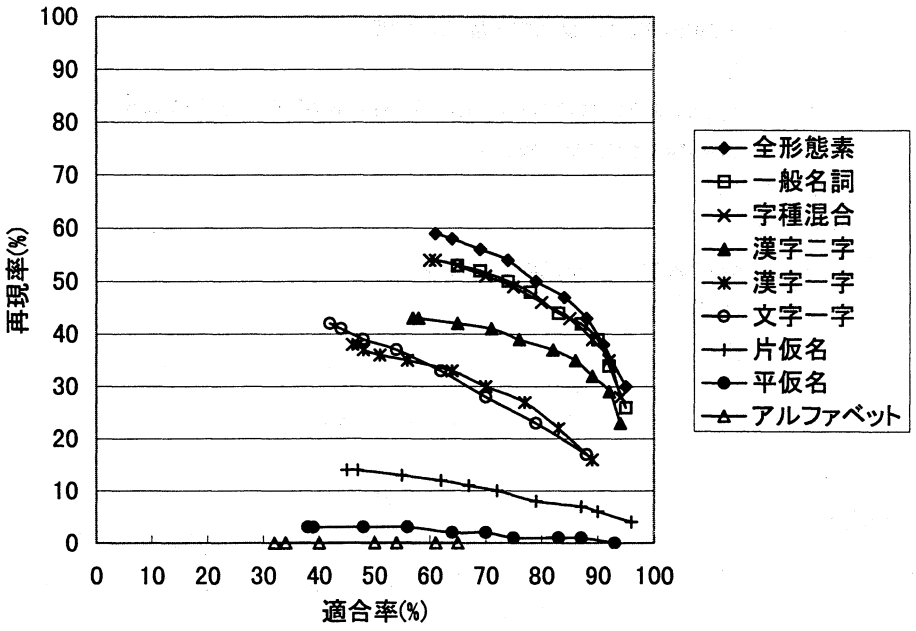
漢字、平仮名、片仮名、アルファベットなど、字種毎にまとめることによって文字列を分解する。ただし、漢字については、さらに2文字づつに分解する[1]。この場合、個々の字種のみを用いる方法と、組み合わせて用いる方法を試した。

2.1.3. 形態素解析による分解

日本語の形態素解析プログラムを用いて分解する。この場合、分解された形態素を無差別に用いる方法と、普通名詞のみを用いる方法を試した。

---

[1] これは、漢字のまとまりの異なり数が莫大なためである。漢字の場合、2文字で意味を成す場合が多いため、2文字に分解することにした。



グラフ 1：職種推定システムの適合率と再現率

2.2. 確率データの準備

実験の準備として、法人顧客データベースから名称と職種を抽出した[2]。この中から無作為に1万件のデータを評価用として分離し、残ったデータを用いて、以下のように文字列が表す職種の確率を算出した。

まず名称から、職種推定に用いるのと同じ方法で文字列を抽出する。次に、文字列 A を抽出した名称のうち、職種が B であるものの数を数え、これを X とする。また、文字列 A を抽出した名称の数を Y とする。この時、文字列 A が職種 B を表す確率を $\frac{X}{Y}$ とする。

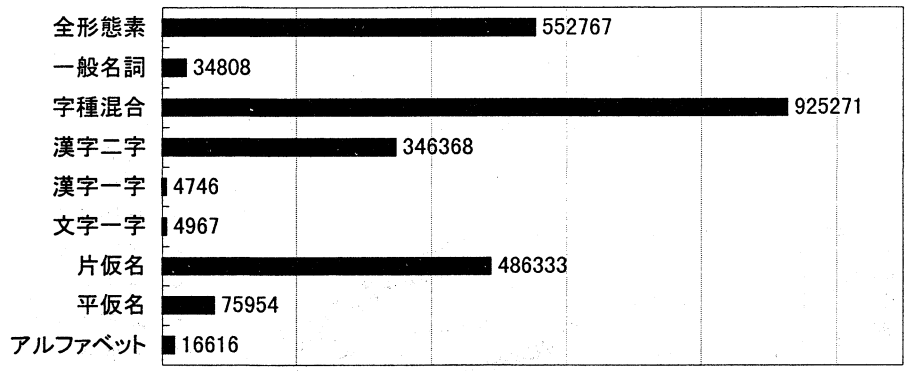
2.3. 職種の推定方法

以下に職種の推定方法を記述する。名称 C を表す文字列を S_0 とし、ここから文字列 $S_1 \sim S_n$ が抽出されたとする。文字列 S が職種 K を表す確率を $P(S,K)$ と表すとすると、名称 C が職種 K である確率 $P(S_0,K)$ は以下の式で計算する。

$$P(S_0,K) = 1 - \prod_{i=1}^{n}(1 - P(S_i,K))$$

この式を用いて全ての職種に対する確率を計算し、その中の最も高い確率が閾値を超えた場合、その職種を名称 C の推定職種とする。この閾値を変化させると、適合率と再現率が変化するため、その推移を実験で計測する。

[2] 約１４００万件



グラフ 2：抽出文字列の異なり数

2.4. 評価方法

評価の対象は、準備段階で分離しておいた1万件の評価用のデータである。データ中の名称に対して職種の推定を行い、データ中の職種と一致するかどうかを調べた。実験では、それぞれの文字列抽出方法に対して、閾値を0から0.9まで0.1刻みで変化させ、適合率と再現率の推移を計測した。

3. 実験結果

グラフ1とグラフ2に、職種推定に用いる文字列の違いに応じたシステムの性能を示す。グラフ1は適合率と再現率の推移を示している。また、グラフ2は、用いた文字列の異なり数を示している。ここで、「全形態素」とは、全ての形態素を用いた場合、「一般名詞」とは、固有名詞を含まない一般名詞のみを用いた場合である。「漢字二字」とは、漢字の2文字のみを用いた場合であり、「漢字一字」とは、漢字の1文字のみを用いた場合である。「文字一字」とは、全ての文字の1文字のみを用いた場合であり、「片仮名列」とは、連続した片仮名のみを用いた場合、「平仮名列」とは、連続した平仮名のみを用いた場合である。「字種混合」とは、平仮名列、片仮名列、アルファベット列、および、漢字の2文字を用いた場合である。

グラフ1を見ると、「全形態素」を用いた場合が最も精度が高い。また、「一般名詞」と「字種混合」はほぼ同じ精度である。「字種混合」で使われる漢字2文字、平仮名列、片仮名列、アルファベット列を個別に見ると、「字種混合」の精度に寄与しているのはほとんど漢字2文字であり、次に片仮名列であることが分かる。平仮名列は職種の推定にほとんど寄与していない。また、漢字1文字だけを用いた場合でも、ある程度の精度を保っており、職種によって名称に使用される漢字に偏りがあることが分かる。

グラフ2を見ると、「字種混合」が最も異なり数が多い。このことは、職種推定を行う際に最も多くのリソースを消費することを示している。但し、「全形態素」と「一般名詞」においては、形態素解析プログラムを使用するため、これに

要するリソースを考慮するべきである[3]。

4. 考察

実験から、企業の名称の4割については、9割以上の正確さで職種を推定することが可能であることが分った。4割という数値は、このシステムのみで企業名と職種の結び付けを行うには不十分であるが、企業名のデータベースと組み合わせて、データベースに登録されていない企業名から職種を推定する方法として有効であると考えられる。

精度は、形態素解析を使って全ての形態素を用いて推定する方法が最も高い。また、形態素解析を使わない方法でも、同水準の精度を出せることも分かったが、これには多くの記憶容量が必要である。実験では、一般名詞のみを使ってもほぼ同じ精度であることが示されている。従って、記憶容量に制限がある場合には、形態素解析を使って一般名詞のみを用いる方法が有効である。

5. むすび

企業の名称から職種を推定するプログラムを試作した。実験の結果、日本語形態素解析システムを用いて名称を分解し、それぞれの形態素が表す職種の確率からその名称の職種を推定する方法が最も精度が高く、約40%の名称について90%以上の正解率で職種を推定することができた。また、同様に普通名詞のみを用いて職種を推定する方法では、より少ない記憶容量で、それほど遜色のない精度で推定できることが分かった。いずれの場合でも、形態素解析を用いることで、精度向上あるいはリソースの節約ができている。さらに、形態素解析を用いない場合でも、文字の組み合わせから職種を推定することが可能であった。特に漢字2文字が職種の推定に有効であった。これらの実験により、企業の名称自身から職種をどの程度推定できるかを示すことができた。

[3] 実験に用いた形態素解析システムの登録語数は約35万語である。グラフ2で示される全形態素の異なり数が35万よりも多いのは、未登録語を名詞として形態素に含めているためである。

参考文献

[1] 岩瀬成人, 大山実, "自然言語処理技術を用いた職業別電話帳検索の高度化," 電子情報通信学会論文誌 D-II, Vol.J74, No.9, pp.1255-1263, 1991.

[2] 渕武志, 米澤明憲, "日本語形態素解析システムのための形態素文法," 自然言語処理, Vol.2, No.4, pp.37-65, 1995.

[3] 渕武志, 松岡浩司, 高木伸一郎, "保守性を考慮した日本語形態素解析システム," 情報処理学会研究報告, SIG-NL97-4, pp.59-66, 1997.