# 交通オントロジーの半自動拡張のための交通用語認識

# Transportation Terminology Recognition for Semi-automatic Traffic Ontology Expansion

河辺一仁　三輪誠　佐々木裕
Kazuhito Kawabe　Makoto Miwa　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute
{sd13409,　yutaka.sasaki,　makoto-miwa}@toyota-ti.ac.jp

## １．はじめに

近年，自動車の自動走行に関する研究が進んでいる．自動車が公道を走るためにはただ道路形状に沿って走るだけでなく，交通法規や交通マナーを守り事故をおこさないようにする必要がある．また，現状の自律走行システムは，通常のプログラムコードの中に走行に必要な知識が組み込まれており，交通法規改正に伴う運行制御の更新が複雑になる．

本研究では，このような背景をもとに走行に必要な交通法規や交通マナーなど，交通に関する必要な知識を独立したオントロジーとして構築することを提案してきた[1]．オントロジーを構築することで，交通法規に関する更新があった場合，自動走行の運行・制御システムを変更することなくオントロジーを書き換えるだけで更新ができるようになる．さらに，構築したオントロジーから車が自動走行する際に必要な情報を取り出せるようにすることを目標としている．

本論文では，大量の交通に関する文書からオントロジーを構築することは困難であるのでオントロジーの半自動拡張を提案する．交通用語や関係性をコンピュータで自動的に抽出が，自然言語処理におけるこのような抽出精度を１００％にすることは非常に困難である．よって，人手を加えることで間違えを訂正し，オントロジーに追加していくことを半自動拡張とする．

## ２．オントロジーとは

オントロジーとは，ある領域の知識を記述，表現するために使用される語彙を定義するものであり，情報の共有が必要な人，データベース，アプリケーションシステムなどによって使われるものである．オントロジーは，コンピュータによって再利用可能なその領域の基本的な概念とそれらの関係の定義を含むものである[2]．本研究では，交通に関する知識を表現するために，交通に関する知識の領域にある語彙を定義し，それら語彙間の関係を定義するものである．

## ３．関連研究

関連研究として，オントロジー研究の基礎と応用を溝口[3]が行っている．また，辞書を用いることによるオントロジーの自動生成の研究について鈴木[4]が行っている．オントロジーを用いた Q&A システムの研究については宮崎ら[5]が行っており，オントロジーからの情報抽出を情報抽出ルールやテンプレートを用いずオントロジー上の活性伝播により情

報抽出を行う研究を廣田ら[6]が行っている．

## ４．提案手法

オントロジーの半自動拡張を行うための順序としてまず，文書データから交通用語の抽出を行う．次に，抽出した交通用語間の関係性を抽出する．抽出したものについて誤りがあるものに対して手動で訂正を行い，オントロジーを構築していく．本稿では，交通用語に対してオントロジーのノードを認識する実験について記す．

## ５．実験

文書データから交通用語抽出を機械学習手法である Conditional Random Fields (CRF)[7][8]を用いて行う．

CRF とは，分類問題における系列ラベリング（入力系列$x$が与えられたときに適切なラベル列$y$を与える）の１種で対数線形モデルを適用したものである．本研究では入力系列を文，ラベル列をカテゴリ別の IOB2 タグとして用いた．

文書データとして，交通教則文を用いた．交通教則文に出現する交通用語をカテゴリ別に階層関係をとり，人手でタグ付与を行った．以下に各データの数を示す．

表１．実験に用いたデータの数

| 交通教則文数 | 総カテゴリ数 | 交通用語異なり数 | 総交通用語数 |
|---|---|---|---|
| 2,940 文 | 57 個 | 842 個 | 4,830 個 |

以下にオントロジーの構造と文書データ中に出現した用語数をかっこ内で示す．

-Animal (3)
-Human (471)
  -HumanParts (54)
-Vehicle (6)
  -Automobile (541)
    -StandardAutomobile (33)
    -SmallSpecialVehicle (8)
    -MediumSizeVehicle (17)
    -LargeTruck (2)
    -LargeVehicle (18)
    -LargeSpecialVehicle (5)
    -Bus (28)
    -EmergencyVehicle (19)
    -Truck (5)
    -ConveyPassengersCar (33)
  -MoterCycle (49)
    -SmallMoterCycle (3)
    -MotorizedBicycle (40)
    -LargeMoterCycle (19)
  -LightVehicle (73)
    -Bicycle (61)
  -Train (2)
-License (29)
  -LargeMotorVehicleLicense (3)
  -LargeVehicleLicense (2)
  -LargeSpecialVehicleLicense (1)
  -MediumSizeVehicleLicense (1)
  -MopedBicycleLicense (1)
  -MotorVehicleLicense (1)
  -SecondClassLicense (2)
  -FirstClassLicense (3)
  -SmallSpecialVehicleLicense (1)
  -TemporaryLicense (3)
  -TractionLicense (3)
  -StandardAutomobileLicense (2)
-Unit (89)
  -Distance (25)
  -Weight (12)
-CarOperate (6)
  -GearState (29)

-Device (201)
  -BicycleParts (36)
  -MoterCycleParts (71)
  -CarParts (506)
-Certificate (8)
-Color (386)
-Illegal (14)
-Limit (183)
-Motion (555)
-PenalRegulation (7)
-Phenomenon (6)
-Place (63)
-Road (798)
-RoadSign (203)
-SpeedProfile (41)
-TrafficRestriction (13)
-Weather (36)

CRF において交通用語抽出に用いた特徴は，前後２単語の範囲にあらわれる Unigram と Bigram と前後２単語の範囲にあらわれる品詞の Unigram と Bigram と Trigram である．

交通教則文を５分割し，１つをテストデータ，残りを訓練データとした５分割交差検定を行う．テストデータの平均文数は 588 文であり，訓練データの平均文数は 2,352 文となった．

## ６．評価

交通用語抽出精度の評価指標として，適合率を再現率の調和平均である F 値を用いる．

$$\text{F 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}}$$

以下に，システムの予測結果と真の結果の正誤表を示す．

表２．正誤表

| | | 真の結果 | |
|---|---|---|---|
| | | 正 | 負 |
| 予測結果 | 正 | TP | FP |
| | 負 | FN | TN |

ここで，適合率とはシステムが交通用語と判定したもののうち，実際に交通用語であった割合である．

$$\text{適合率} = \frac{TP}{TP + FP}$$

再現率とは文書内の交通用語のうち，システムが交通用語であると予測できた割合である．

$$\text{再現率} = \frac{TP}{TP + FN}$$

実験から，すべての交通用語に対する抽出精度は下記の表３のように得られた．各カテゴリにおけるそれぞれの抽出精度は次ページの図１に示す．

表３．全体の抽出精度

| 再現率 | 適合率 | F 値 |
|---|---|---|
| 0.763 | 0.883 | 0.819 |

## ７．まとめと今後の課題

交通教則文にカテゴリ別にタグ付与を行い，CRF を用いた用語抽出を行った結果の抽出精度として，約８割の精度を得ることができた．半自動拡張を目的とした精度と考えると高い精度を得られたのではないかと考えられる．

用語抽出において失敗した例として，文書中で１，２回のみの出現であるもの，交通用語のカテゴリが曖昧であるもの（車の後部座席とバイクの後部座席など）があげられる．用語抽出の精度をあげるためにこれらの問題を解決することが今後の課題であると考える．

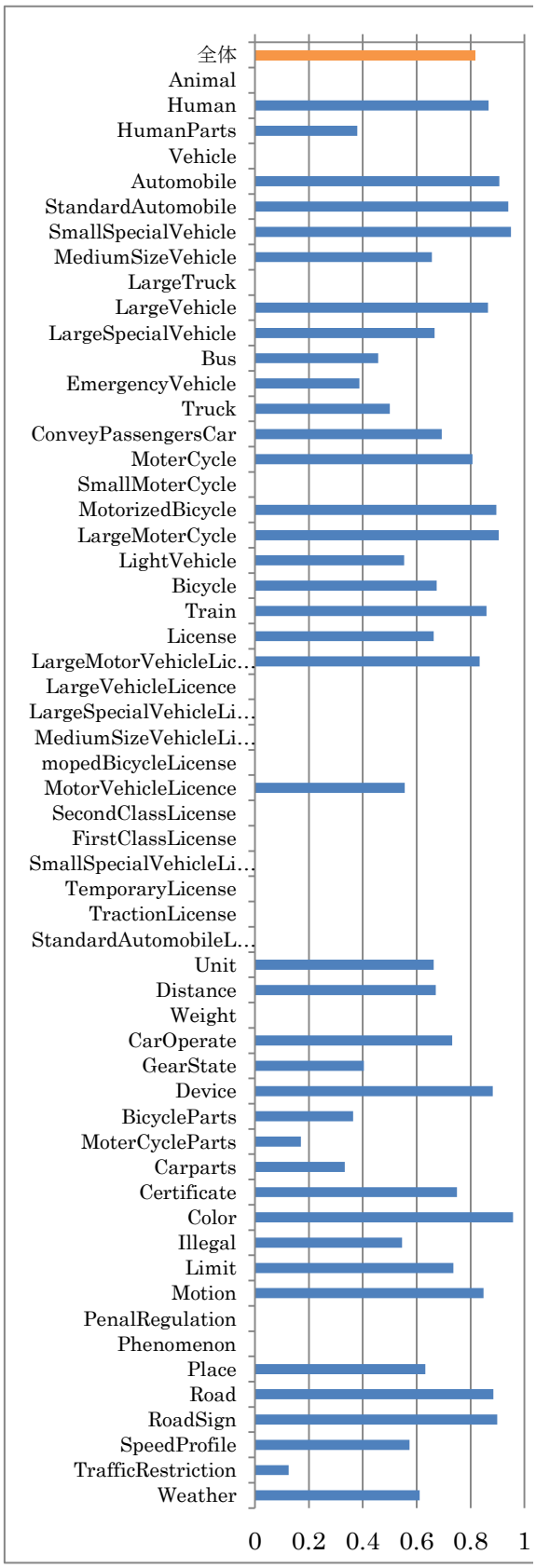


図１．各カテゴリの F 値

具体的な手法として，カテゴリの曖昧性を解消できるような CRF の特徴を新しく考案すること，出現頻度の低いカテゴリの用語を必ず訓練データで学習できるようにするなどが考えられる．

また，交通用語間の関係性抽出も行っていく．オントロジーに抽出した用語を追加しそれらに付随する性質などを付与し，それらの情報を用いた関係性抽出を行っていく．

最終的には，構築できたオントロジーの評価を情報抽出として，交通法規問題に対する Q&A システムの作成を目指す．

参考文献

[1] 杉村皓太：交通法規問題の解答システムの向上, 2013

[2] 小林一郎：人工知能の基礎，2008 pp.72-83.

[3]溝口理一郎：オントロジー研究の基礎と応用，1999，人工知能学会誌 Vol.14 No.6 pp.977-988

[4]鈴木敏：辞書からの上位五情報抽出とオントロジー自動生成，2009 自然言語処理 Vol.16 No.1 pp.101-116.

[5]宮崎勝 他: Q&A システムのための野球オントロジーの設計に関する検討，2005 2005 年映像情報メディア学会冬季大会 成蹊大学

[6]廣田啓一 他: オントロジー主導による情報抽出，1999，人工知能学会誌 Vol.14 No.6 pp.1010-1018

[7]DANIEL JURAFSKY & JAMES H. MARTIN：SPEECH AND LANGUAGE PROCESSING , 2008, pp.235-241

[8]高村大也：言語処理のための機械学習入門, 2010, pp.132-159.