3D3-1

移動要求抽出のための移動目的発言分類法に関する一考察

A Study on Traveling Purpose Classification Method to Extract Traveling Requests

鈴木信雄[*1]
Nobuo Suzuki

津田和彦[*2]
Kazuhiko Tsuda

[*1] KDDI 株式会社
KDDI Corporation

[*2] 筑波大学
University of Tsukuba

It is possible to get the flow information of people and transportation efficiently by collecting travelling information in the Internet. In the meantime, Q&A sites on the Internet express human requests more directly and they are the useful resources to extract many kinds of knowledge and intentions. This paper proposes the classification method for speeches in Q&A sites based on the road traffic census by MLIT in Japan to extract traveling requests. Specifically, this method presumes the traveling purposes by SVM. Learning with the frequency of parts of speech that express traveling requests and TFIDF as the features of SVM is carried out. We also evaluated the performance by the experiment and the accuracy was 45.5%.

1. はじめに

現在, 道路の混雑状況などの情報は, 道路に設置されたセンサーや個別の車両から収集される位置情報などにより生成されている. しかし, これらの情報だけでは, 人間や鉄道などの流れを考慮したグローバルな混雑状態を把握することは難しい. これに対して, インターネット上で公開されている情報の中には, 地理的な移動に関する要求の情報が大量に存在する. これらの情報を抽出し, 人間や車両などの移動情報を収集することで, 交通の流れを効果的に把握できる可能性がある. 一方, インターネット上で広く利用されている質問応答サイトは, 人間の要求の情報が直接的に良く表現されており, 各種の知識や意図の抽出に有用なインターネット上のリソースである. したがって, 本研究では, このような質問応答サイトの中に出現する移動要求情報を抽出し, 人間の移動状況を収集することで, 交通の流れを把握することを目標としている.

本稿では, この目標を実現するために必要な質問応答サイトからの移動要求情報の抽出手法について提案する. 具体的には, 国土交通省における道路交通センサスの移動目的分類をベースとした質問回答サイト内の発言単位での移動目的分類法を検討した. 本手法では, SVM を用いて分類を行い, 素性としては, 形態素の頻度順位, および, 移動要求に特徴的に現れる品詞の TFIDF を用いている. 次項より, 分類手法の内容を示すと共に, 評価実験を行った結果とその分析について述べる.

2. 移動目的の分類

質問応答サイトにおける移動要求発言の例を表 1 に示す. それぞれの発言からもわかるように, このような発言からは, 移動の目的, 移動する人の年齢・性別・職業, 移動人数, 出発地, 目的地, 移動手段, 移動ルート, 移動要求の強さなどの情報を抽出することができると考えられる. これらの中で, 今回は, 移動要求の最も基本的な情報である移動の目的に着目した.

移動の目的を分類するためには, 参照する分類の定義が必要である. 様々な分類が考案されているが, 本研究では, 交通の移動を主に扱うことから, 国土交通省が行っている「交通センサス」にて定義されている分類を利用することとした[国交省 09].

連絡先：鈴木信雄, KDDI 株式会社, nu-suzuki@kddi.com

表 1. 質問応答サイトにおける移動要求の発言例

9月に親子3代で富士山麓～箱根～伊豆方面へ旅行予定です. 箱根（芦ノ湖周辺）から伊豆・北川温泉まで自家用車での移動ですと, どのくらいの時間がかかるのか教えてください.
中部国際空港から妻籠まで行こうと思っています. 名古屋の首都高速をとおり, 小牧ICから中央道で行くのが早いのか, 東海環状自動車道, 中央道が早いのか教えてください. ちなみに中部国際空港は10時前後に出発予定です.

表 2. 交通センサスの移動目的分類表

No.	項目	説明
1	出勤	通勤のため会社へ行く場合.
2	登校	就学先への登校, 校外活動. 塾含まず.
3	家事・買物	業務での買物は含まない.
4	食事・社交・娯楽	日常生活圏内の私的なつきあい, 映画等
5	観光	観光名所, 旧跡などへの観光.
6	保養	温泉, 家族・知人との交流などの保養.
7	スポーツ	ハイキング, ゴルフ, 運動会などスポーツ.
8	体験型レジャー	遊園地・ドライブ・名産品の飲食等.
9	その他私用	通院・習い事など.
10	送迎	送迎（業務での送迎は含まない）.
11	荷物の運搬を伴わない業務	業務目的で荷物を運搬しない場合
12	荷物の運搬を伴う業務	業務目的で車を利用した場合.
13	帰社	業務が終わって会社へ戻るための運行.
14	帰宅	勤務先通学先, 買物, 外出先から自宅.
15	その他	上記以外のその他.
16	不明	移動目的が不明な場合.

この交通センサス移動目的分類を表 2 に示す. ここで, 発言中に移動目的を含んでいない場合には,「不明」に分類した.

3. 素性の選択

前項で述べた移動要求発言では, 明確に移動の目的が表現されていない場合が多い. 例えば,「伊豆方面へ旅行予定です」とあれば, 温泉旅行であることは明示されていなくとも, 表 2 の移動目的分類における「保養」である可能性が高い. そのため, 明示的な単語辞書の構築による分類を行うのではなく, 近年注目されている SVM を使った分類を行うこととした. また, SVM の素性には以下の 2 種類のデータを試みた.

① 各分類における品詞の頻度順位

学習用に収集した 225 個の移動要求を含む発言における品詞の頻度割合を解析すると，表 3 のような結果を得た．移動要求発言においては、これらの品詞が常に高頻度で出現すると考えられるが，その頻度の順位は，発言によって異なっている．例えば，「保養」に分類されるような発言であれば，地名やホテル名が上位に出現し，「通勤」であれば，時刻や移動手段が上位となる．そこで，各移動目的分類における特定の品詞に対する形態素の頻度順位を SVM の素性として用いることとした．具体的な品詞は，表 3 に示す 10 個の品詞を用いた．

表 3. 移動要求発言中の移動目的関連品詞の頻度

品詞	割合	単語の例
動詞-自立	10.1%	着く，行く，帰る
名詞-一般	10.1%	タクシー，新幹線，高速，ホテル
助詞-格助詞-一般	9.5%	から，に
名詞-固有名詞-地域-一般	4.6%	品川，表参道
名詞-サ変接続	4.4%	出発，到着，迂回，渋滞
未知語	3.0%	ＥＴＣ，１６０ｋｍ，アクアライン
名詞-副詞可能	2.3%	火曜日，朝、普段
名詞-固有名詞-一般	0.7%	東名高速道路，ディズニーランド
名詞-固有名詞-人名-姓	0.4%	三島，豊科，熊本
名詞-固有名詞-組織	0.3%	大磯プリンスホテル，北里大学

② 特徴的に出現する品詞における TFIDF

TFIDF は，一般に文書に対する代表キーワードを抽出するための指標として多く用いられている．移動要求発言においても，代表キーワードを使えば，高精度の分類を行えると期待できる．ただし，単なる単語の TFIDF を求めるだけでは，助詞などの雑音が多く，精度低下の原因となる．そこで，移動目的を含む発言の特徴を表現している「名詞-一般」と「名詞-サ変接続」を対象として，品詞毎の形態素 TFIDF を素性として用いることとした．

ここで，品詞毎の形態素 TFIDF は以下の式のように表すことができる．tf は 1 つの発言中に含まれる当該品詞の形態素頻度，df は移動要求発言の中で当該品詞の形態素が含まれる発言数，N は学習用の総発言数である．

$$tfidf_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i})$$

4. 評価実験

実際の質問応答サイトのデータを使って，これまで述べた分類手法の評価実験を行った．評価実験用データとして 450 個の発言を収集した．この中で，学習用として 225 発言，評価用として 225 発言を使用した．学習用では，正例が 68%，負例が 32%，評価用では，正例が 84%，負例が 16%であった．また，SVM ツールとしては，TinySVM を利用した[TinySVM]．カーネル関数は TinySVM のデフォルトであるリニアを用い，コストマージンパラメータ C もデフォルトの 1.0 とした．

実験は，まず，学習用データから発言毎に 3 項で述べた 2 種類の素性を求めモデルを構築した．具体的には，収集した学習用データの形態素を茶筌により求め，発言単位の形態素頻度を算出した．次に，特定品詞における形態素毎の TFIDF の値を求めた．この 2 つの値を発言毎に学習しモデルを作成した．

つづいて，このモデルを使って評価用データの発言毎に推定を行った．評価は，各発言を人によって判別した分類と，この推定結果とを比較して同一であれば正解とした．実験の結果，正解率は 45.5%となった．正解と不正解の例を表 4 に示す．

表 4. 正解と不正解の例

	発言内容	推定	正解
正解例	今週末か来週末にぶどう狩りに行こうと思っているのですが，福岡県もしくは大分県のぶどう園で，食べ放題ができるぶどう園を教えてください．	体験型レジャー	体験型レジャー
不正解例	来週，初めて大阪に出張することになりました．レンタカーを借りて堺市と東大阪市を回る予定です．翌朝レンタカーを堺駅前で借りようと思っているのですが，大阪駅に行ったほうがよいのでしょうか．	出勤	荷物の運搬を伴わない業務

表 5. 不正解データの分類

No.	分類	割合
1	特徴的な形態素が存在するのに不正解	52.8%
2	特徴的な品詞が存在するのに不正解	33.7%
3	移動目的表現が無いのに不適切な移動目的に分類	13.5%

不正解のデータを解析すると，表 5 のように 3 種類のデータに分類できることがわかった。まず，「特徴的な形態素が存在するのに不正解」では，移動目的を表す特徴的な形態素が正例の学習データ中に含まれていないために不正解となっていると考えられる．これらの発言は，現状の素性で定義された表 3 の品詞データにより移動目的が推定できると考えられるが，特徴的な形態素が正例の学習データ中に含まれていないために，不正解となっていると考えられる．例えば，「ライブ」という形態素から「食事･社交･娯楽」の分類が推定できるものなどがある．これに対しては，正例学習データを増やすことにより，正しい推定が可能と思われる．次に，「特徴的な品詞が存在するのに不正解」では，TFIDF を求める移動目的を表す特徴的な品詞が含まれていないために不正解となっていると考えられる．例えば、「食べられる所」の「動詞-自立」などがある．これに対しては，TFIDF を求める品詞に対して，次の品詞を加えることで正しい推定が可能と思われる．「名詞-固有名詞」，「名詞-接尾」，「名詞-非自立」，「動詞-自立」，「動詞-接尾」．最後に，「移動目的が無いのに不適切な移動目的」は，そもそも移動目的が存在しないにもかかわらず，不適切な移動目的に分類されたものである．これは，負例の学習データが不足しているために不正解となったと考えられる．これらの発言を見ると，いずれかの分類に当てはまるような特徴的な形態素が散見されるが，全体としては，移動目的を示していない発言となっている．これに対しては，負例のデータを増加し学習することで「不明」の分類に推定可能と思われる．

5. まとめ

本稿では，移動要求を含む質問応答サイトの発言に対して，移動目的による分類を行う手法を提案した．この手法では，SVM を利用し，品詞の頻度順位と特定品詞における TFIDF を素性として用いた．実験の結果，十分な性能が得られるとは言えなかったが，不正解データの解析を行い，学習データ数が不十分であることや，素性で利用した品詞の種類の妥当性などが原因と思われることがわかった．今後は，これらの考察を具体的にシステムに組み込んで再度評価を行う予定である．

参考文献

[国交省 09] 国交省：道路交通センサス，
http://www.mlit.go.jp/road/census/h21/index.html
[TinySVM] http://chasen.org/~taku/software/TinySVM