ブログ記事における Entity Linking を利用したスポット抽出

粟村 誉　工藤 和也　井上 裁都　宮崎 林太郎　山下 達雄
ヤフー株式会社
{tawamura, kkudou, tatinoue, rimiyaza, tayamash}@yahoo-corp.jp

1　はじめに

近年，ユーザ投稿型サービスにおいてコンテンツにリッチな店舗・観光名所情報を付与するニーズが高まっている．著者らは Entity Linking[1] を利用した，ブログ記事投稿サービス（Yahoo!ブログ[1]）の改善を行っている．ユーザの投稿した記事に店舗や観光名所（以降，これらを**スポット**と呼ぶ）に関する言及がある場合，Yahoo!ロコ[2]などに含まれるスポット情報と紐づけることで，サービス間回遊の円滑化が期待できる．これについて，テキスト解析技術を用いたレコメンドにより，ユーザ自身によるスポット情報登録の支援が可能である．本稿では，ブログ記事投稿サービスにおいて，ユーザの投稿した記事のテキストからその記事に関するスポットの推薦を行うシステム（図 1）について述べ，その評価を行う．

2　ブログ記事からのスポット抽出

本稿で扱うスポットは，Yahoo!ロコに登録されているものとする．Yahoo!ロコでは，検索可能なスポットについての情報が表 1 のように登録されている．本稿では，ブログ記事から抽出した情報をもとにスポットの検索をすることで，適切なスポットを推定するシステムを提案する．その手法について以下で述べる．関連研究として佐々木ら [2] の報告などが挙げられる．

2.1　スポット抽出に使用する情報

以下の 4 つの情報をブログ記事より抽出し，Yahoo!ロコでの検索に使用する．以降，下記情報の抽出対象は各ブログ記事のタイトルとサブタイトル，記事本文を連結させた一つのテキストとする．

2.1.1　タグ

記事中に含まれる特徴語を記事のタグとして抽出する．特徴語抽出は形態素解析を行い，ルールベースに

[1]https://blog.yahoo.co.jp
[2]https://loco.yahoo.co.jp

図 1: 記事の投稿確認画面におけるスポット推薦例

よる形態素の連結処理を行うことで生成する．また，連結した特徴語について，それぞれスコアを以下の式から算出する．

$$score = TF \times IDF \times \text{文字列長} \quad (1)$$

ここで TF は入力記事中における対象特徴語の頻度である．IDF は事前に大規模なコーパスを使用して算出した形態素の idf 値を使用し，各特徴語についてそれぞれの構成する形態素の idf 値の平均値をその特徴語の IDF として扱う．また 2.1.2 節で抽出できる「東京都港区」のような地域情報にあたるものはタグの候補から外す．本稿では抽出したタグのうち，システムへの負荷を考慮しスコアの高いものから最大 5 つのみを使用する．

2.1.2　地域情報

記事に含まれるランドマークや地名から記事の言及している地域を推定し，抽出する．このようにして得られる記事の代表的な地域を，記事の地域情報と呼ぶ．

まず Entity Linking によって記事からエンティティを抽出する．次に抽出したエンティティが地域・住所情報を含む場合，どの地域についてのエンティティが

表 1: Yahoo!ロコ上の検索可能な情報例

種類	例
店名	つるとんたん 六本木店
緯度経度	139.7343**, 35.6625**
カテゴリ	しゃぶしゃぶ、うどん、居酒屋、懐石・会席料理、懐石（懐石料理）、会席料理、日本料理、和食、そば・うどん
読み	ツルトンタン
住所	東京都港区六本木 3-**-**
別表記	つるとんたん琴しょう楼、麺匠の心つくし つるとんたん 六本木店
電話番号	03-5786-****
エリアコード	13, 13103, 13103029, 13103029003, 1310302900300014

多いかをスコアリングすることで，地域情報を推定する．例えば，記事から「東京タワー」や「港区 (東京)」などのエンティティが得られた場合,「東京都港区」の地域に関する記事である可能性が高いと推定できる．

推定した各地域のスコアは，より信頼性の高い地域の選択に使用できるほか，確信度の高い推定地域のみ扱うための閾値としても使用できる．Entity Linking の具体的な手法については石川ら [3] の報告を，地域情報推定の具体的な内容については長田ら [4]，井上ら [5] の報告を参照されたい．

2.1.3 POI

スポットについての記事の中には POI（住所文字列）が含まれるものもある．このような POI を抽出し，緯度経度に変換することで言及されているスポットを推定する．

POI には記事によって表記揺れが存在するため，本稿では，2.1.1 節で抽出したタグの中から，住所文字列になっていそうなものを POI として抽出する．抽出は以下のロジックで行う．

1. 「愛媛県伊予郡松前町大字東古泉 50 番地 100」などの文字列が，タグ抽出によりタグの候補として得られる
2. 文字列に「愛媛県伊予郡松前町」などの地名エンティティへのメンションが含まれるか判定する
3. 更に以下の 2 条件を満たす場合，POI とする
   (a) 「郡，区，町，村」のいずれかを含む
   (b) 数字を含む

また，POI として抽出された文字列はタグから除外する．

2.1.4 電話番号

記事によってはスポットの連絡先が含まれる場合がある．Yahoo!ロコではスポットに電話番号情報が付与されている場合があるため，抽出した電話番号から直接スポットを推定することができる．電話番号の抽出は正規表現によって行う．

2.2 アドホックなタグの追加

2.1.1 節のタグは，スポット名となるような特徴的な単語を抽出できるが，一般名詞や助詞などからなる自然文のようなスポット名は抽出しにくい．本稿では，この 5 つのタグに加えて以下のアドホックなルールを新たに適用してタグを追加し，精度の比較を行う．

1. 「〜」『〜』【〜】部分を抽出
   例．宮城県登米市、「うなぎの東海亭」で、うな重二段重ね 3,600 円。
2. 会話文などは対象外とする
   （byte 数制限，感嘆詞で終わるもの削除）
3. 抽出したもののうち，頻度の高い 2 つに絞り，タグへ追加

2.3 スポット抽出とランキング

2.1 節で説明した 4 つの情報を使用してスポットを推定する手法について説明する．

まず，タグと地域情報について述べる．3 章でも述べるが，これらは単体でのスポット推定精度が期待できない．そのため，本稿ではこの 2 つの情報を組み合わせることによって精度の高い推定を実現する．具体的には，地域情報推定により記事が言及している地域を推定し，その地域についてタグでスポットの検索を行うことで，該当地域内に限ったスポットへ絞り込みを行う．タグは各ブログ記事について複数持つことを想定しており，各タグから得られた検索結果を統合・スコアリングすることでより適切なスポットを推薦できるようランキング処理を行う．スコアリング手法は以下の通りである．

1. 各タグについて Yahoo!ロコから得られた 10 件のスポットはすべて+1pt
2. スポット名とタグの一致による加点
   - 完全一致：+5pt
   - 7 割以上一致：+4pt
   - 5 割以上一致：+2pt
   - 部分一致：+1pt

表 2: スポットのスコアリング例

地域：東京都港区六本木	
タグ	検索結果
つるとんたん	つるとんたん 六本木店 (+1, 一致+2)
うどん	つるとんたん 六本木店 (+1) 丸亀製麺 六本木ティキューブ店 (+1) 福の膳 六本木店 (+1)
春キャベツ	-

表 3: ブログ記事 100 万件に対する抽出可能記事数

抽出する情報	抽出可能記事数
タグ	983647
地域情報 (閾値以上)	267457
地域情報 (閾値なし)	418667
POI	18012
電話番号	15732

例えば，表 2 のように，地域情報が「東京都港区六本木」でタグが “つるとんたん”，“うどん”，“春キャベツ” だった場合,「つるとんたん 六本木店」は 4pt，それ以外のスポットは 1pt となり，スコアの高い「つるとんたん 六本木店」が優先される．

タグ・地域情報と異なり，電話番号と POI については単体で推定を行う．電話番号はそのままクエリとして Yahoo!ロコを検索し，得られたスポットが推定結果となる．POI については，まず POI を緯度経度に変換し，その座標から半径 10m 以内のスポットのみを Yahoo!ロコから取得する．複数得られる場合，タグを含まない記事では近い順に，タグを含む記事では前述したタグでの検索のスコアリングも併用し，もっともスコアの高いスポットが選択される．

3 実験と評価

2 章で述べた手法でスポット推定を行った．ブログ記事データとして，2012〜2016 年に作成されたブログ記事 100 万件を使用する．

推定評価を行う前に，ブログ記事 100 万件について地域情報や POI などがどのくらい抽出可能かを確認する．抽出結果は表 3 に示す．地域情報は，定性的に地域推定精度が向上していると判断した地域情報スコアの閾値を設け，それぞれでの抽出可能記事数を算出した．このように，タグや地域情報は多くの記事から抽出できるにも関わらず，POI や電話番号が抽出できる記事はいずれも 2%以下である．これは多くの記事においてスポットを含むわけではなく，またスポットを含む場合でも POI や電話番号が詳細に書かれている記事が少ないことを示唆している．

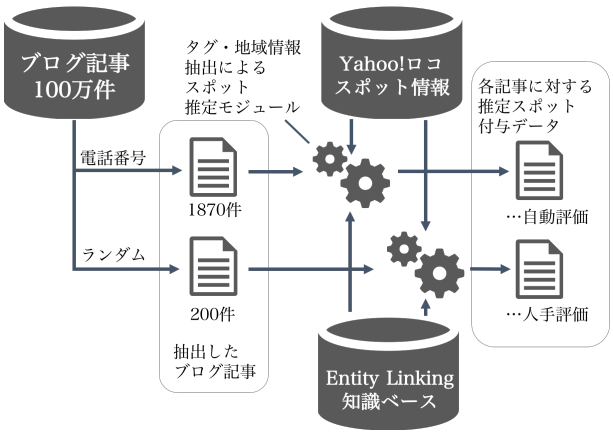

図 2: 使用データと評価実験の概略図

推定内容が実サービスでの利用に耐えうるかを定性的に評価したところ，POI や電話番号はスポット情報を直接指していることが多く，スポット推定精度が高いことがわかった．逆にタグや地域情報はスポットを絞り込むことが難しく，スポット推定精度は低い．よって，本稿ではタグと地域情報を組み合わせることで推定するスポットの精度を向上することを検討する．

評価は，電話番号での推定結果を正解とする自動評価とランダムに選択した記事に対する人手評価の 2 通りについて行う．使用データと評価実験の概略図を図 2 に示す．

3.1 スポット推定実験（自動評価）

電話番号による推定スポットを正しいと仮定して正解データを作成し，そのスポットをタグと地域情報により推定できるかどうかを評価した．電話番号が抽出でき，かつ Yahoo!ロコのいずれかのスポット情報と紐つけることができたブログ記事は 1870 件あった．これらはすべて正解を含む記事となる．

手法として，POI のみによる推定を行うもの，2.1.1 節で述べたタグを 5 つまで使用するもの（Baseline），2.2 節のアドホックなタグを追加したもの（AddTag）の 3 手法を比較した．Baseline と AddTag で用いる地域情報については，表 3 で示した “閾値以上” のものとした．評価はシステムが出力した上位 1 件，3 件にそれぞれ正解が含まれるかで検証した（Pre@1，Pre@3）．また同時に，正解かどうかに関わらず何らかのスポットを推定できた記事数（Extracted），抽出に使用した平均タグ数（Tag）も同時に算出した．

結果を表 4 に示す．表 4 から，Baseline よりアドホックなルールを追加した手法の方が精度が向上し，タグ数の増加も 1 以下に抑えられた．POI は Pre@1 では Baseline に劣っているが，Pre@3 では他の手法

表 4: 自動評価結果

Method	Extracted	Pre@1	Pre@3	Tag
POI	1469	0.78	0.91	-
Baseline	1496	0.79	0.86	5.0
AddTag	1663	**0.84**	**0.89**	5.9

よりも精度は高かった．POI での誤り例として，得られた緯度経度付近に該当するスポットが複数抽出される，記載されている住所がスポットに関するものではない，などが挙げられる．

AddTag での推定誤りの例として，「宇治ミルク金時」のような地名混じりの食品名を含む記事が挙げられる．この記事では “宇治” の部分に「宇治市」エンティティが付与され，誤った地域情報を付与してしまう．これらは別途食品名辞書などを作成することで，食品名として正しく抽出することで回避することが考えられる．

3.2 スポット推定実験（人手評価）

3.1 節での評価は全て電話番号とその関連するスポットを含むブログ記事に対してのみである．これらの情報を含むブログ記事は推定が容易であると言え，また実際のブログ記事にはスポットの推定が不要なものも存在する．そのため，3.1 節の実験に加え，新たにランダムに選択した 200 件のブログ記事に対して，AddTag での推定と，人手での評価を行った．

結果を表 5 に示す．また正解の有無やそれに対する推定結果の頻度を表 6 に示す．ここで G:1，G:0 はそれぞれブログ記事に正解となるスポットが存在する/しないを表し，Mat@1 は Top1 件で正解した記事数，Mat@2,3 は Top2,3 件目で正解した記事数，Unmat は誤ったスポットのみ出力した記事数，No Mat はスポット推定ができなかった記事数を表す．ブログ記事に正解が存在するかは記事のテキストのみから判断し，記事中の画像などのみからしか判別できないものは，正解となるスポットは存在しないものとして判断した．

表 5 の Precision の結果から，3.1 節での電話番号を含む記事に比べ精度が悪化していることがわかる．これは電話番号を含まない記事では，記事に含まれるスポットの情報量が少ないためであると言える．また，Top1, Top3 のいずれにおいても Precision が Recall を下回っている．これは表 6 からわかるように，正解のない記事に対して誤った出力をしている頻度が 45 回と多いためである．

誤った例としては，“雪の小山” という文脈から「栃木県小山市」と誤った地域情報抽出をしてしまったもの

表 5: 人手評価結果

	Top1	Top3
Precision	0.36 (50/140)	0.43 (60/140)
Recall	0.44 (50/113)	0.53 (60/113)
f-measure	0.40	0.47

表 6: 正解とシステム出力の対応頻度

	Mat@1	Mat@2,3	Unmat	No Mat
G:1	50	10	35	18
G:0	-	-	45	42

や，“コスモス” や “ごはん” といった一般的な名称を示すタグから，具体的にそれらの名前を店名に含む飲食店を抽出してしまったものなどが挙げられる．これらについては Entity Linking の精度向上や，確信度の高い推定結果のみ出力する閾値の設定などの対応が考えられる．

4 おわりに

本稿では，ユーザが執筆したブログ記事のテキスト情報から言及しているスポットを推薦するシステムの構築と評価を行った．本システムは公開中の実サービスに導入済みである．

今後の予定として，実際に使用したユーザの推薦項目の使用ログなどから評価と分析，それに対する改善を検討している．また，本稿ではスポット推定の評価のみを行ったが，サービス上ではタグを実際にブログ記事に付与するタグとして推薦しており，これらのタグの妥当性の評価も，使用ログなどをもとに進めたい．

参考文献

[1] Rada Mihalcea and Andras Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. ACM, 2007.

[2] 佐々木彬, 五十嵐祐貴, 渡邉陽太郎, 乾健太郎. 場所参照表現のグラウンディングに向けて. 言語処理学会第 20 回年次大会, March 2014.

[3] 石川裕貴, 小林健, 長田誠也. ウェブ検索ログと Wikipedia 内部リンクを用いたエンティティの曖昧性解消. 言語処理学会第 21 回年次大会. ヤフー株式会社, March 2015.

[4] 長田誠也, 末永圭吾, 善積正伍, 庄司和正, 吉田享晴, 橋本恭明. エンティティリンキングを用いたドキュメントに対する地点情報の付与とその応用. 言語処理学会第 21 回年次大会. ヤフー株式会社, March 2015.

[5] 井上裁都, 末永圭吾, 長田誠也, 立石健二. Entity Linking を用いたニュース記事に対する市区町村単位の地域情報の付与. 言語処理学会第 22 回年次大会. ヤフー株式会社, March 2016.