RI-009

二言語間のタグ変換を用いた画像タグ付与システム

Automatic Image Tagging System with Translation between Two Languages.

石先　広海 †　Hjalmar Olof Wennerstrom ‡　帆足　啓一郎 †　滝嶋　康弘 †
Hiromi Ishizaki　Hjalmar Olof Wennerstrom　Hoashi Keiichiro　Yasuhiro Takishima

1　はじめに

近年のデジタルカメラやカメラ付き携帯電話の普及により，個人が多くのディジタル写真を蓄積する事が可能となっている．また，ソーシャルメディアの急激な発展によって，多くのユーザが個人の写真や映像を WEB 上で共有することが一般的になりつつある．一方で個人が蓄積した写真を目的に応じて検索や閲覧する事は非常に多くの労力を要する．現状では，個人が蓄積した写真や，WEB 写真を検索するためには，人手による写真データの分類やインデキシング作業が必要であり，WEB 上では写真に対して，投稿者もしくは閲覧者が付与したタグなどを利用して写真を検索することが主流である．

本稿では，写真の分類や閲覧に要する労力を軽減させるために，未知の入力画像に自動でタグを付与させるシステムに着目する．さらに，学習データが少ない言語環境においても精度の高いタグ付与システムを構築するために，学習データを WEB 上の画像共有サイトより収集し，タグに多く利用されている英語を用いたタグ付与モデルを構築し，英語タグを日本語タグに変換することで，システムの精度向上を図る．また，提案方式の有効性を検証するために，被験者による主観評価実験を実施する．

2　関連研究

本章では，自動タグ付与システムの従来研究について紹介する．入力画像にタグを付与するために，カテゴリ分類を利用するシステム [1] があげられる．文献 [1] では事前に正解カテゴリ情報が付与された学習データセットを利用することでカテゴリ分類を行う．また，確率モデルを利用する方式として文献 [2] があげられる．画像と画像に付与されたタグ情報間の関係に基づいて確率モデルを構築し，入力画像に対して，最大の確率値を示すタグを付与している．この様に，従来研究ではカテゴリを識別するための識別器や，学習モデルに基づいてタグを付与させている．これらシステムでは，学習データにおける画像と正解情報もしくはタグの量と質に依存して精度が変動するため，学習データの質を確保するための人手による正解情報付与作業が必須となり，多くの労力を要していた．

この様な作業を軽減するために，WEB 上の画像共有サイトなどから，画像と画像に付与されたタグを大量に取得し，学習データとして利用するシステム [3][4] が報告されている．Monay らは，pLSA[5] と呼ばれる潜在トピックを利用した共起確率モデルにより，自動タグ付与を実現しており，学習画像とタグに基づく共起確率 (t-pLSA) と，学習画像と画像特徴量に基づく共起確率 (v-pLSA) を利用してタグ付与モデルを構築している．

具体的には，学習プロセスにおいて学習画像とタグ及び画像特徴量の共起確率から潜在トピックを学習させている．学習画像群が与えられたとき，t-pLSA は潜在トピックを用いて学習画像とタグの共起確率を表現する．この時の潜在トピックに対するタグ及び学習画像の帰属確率を EM アルゴリズムにより計算する．次に，v-pLSA に対して，潜在トピックに対する学習画像の帰属確率を代入し，潜在トピックに対する画像特徴量の帰属確率を計算する．最終的に，未知画像を入力した時に，v-pLSA を利用して潜在トピックに対する入力画像の帰属確率を求め，t-pLSA を適用することで入力画像に対するタグ発生確率を計算し，確率値に基づいてタグを付与している．

3　問題点

この様に，学習ベースの方式においてある程度の精度を得るためには，大量の学習データが必要であり，画像共有サイトから大量の画像とタグを収集することで，人手による学習データ整備作業は軽減される．

しかし，WEB 上に存在する画像共有サイトの多くは，複数の言語によってタグが付与されており，単純に学習データとして画像を大量に取得した場合に，複数の言語が混在した学習データとなることが予想できる．この様な学習データを用いて，自動タグ付与システムを構築した場合に，システム利用者の意図しない言語が結果として付与される可能性がある．例えば，学習データに大量の英語タグと少量の日本語タグが含まれていた場合に，英語のタグが多く付与されていたり，意味のない日本語タグが付与されることなどが予想できる．

学習データに含まれるタグが少数である言語 (小規模言語) がタグとして付与されない問題を改善するためには，小規模言語タグが付与された画像群のみを学習データとして利用する事も考えられるが，学習データに多く含まれる言語 (大規模言語) に比べて少量の学習データとなるため，モデルを構築するのに十分な学習データが確保できず，大規模言語を学習データとしたシステムに比べて精度が劣化する可能性がある．

4　提案システム

本章では，小規模言語を出力する場合においても精度の高い自動タグ付与システムを実現するために，大規模言語と小規模言語が混在する学習データから，ベースとなる大規模言語と出力対象となる小規模言語を抽出し，両言語を学習データとして利用する自動タグ付与システムを提案する．図 1 に提案システムの概要を示す．

本システムは，主に学習データ収集プロセスとタグ付与プロセスに分類できる．学習データ収集プロセスにおいて，WEB 上の画像共有サイトより，学習データとなる

† KDDI 研究所，KDDI R&D Laboratories Inc.
‡ School of Engineering, Uppsala University.

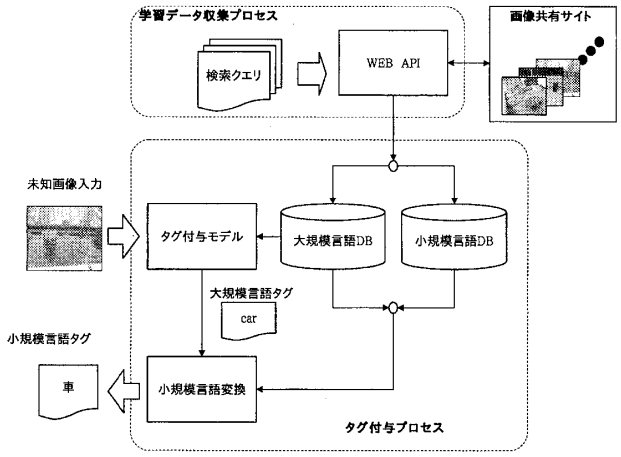

図 1　提案システム概要図

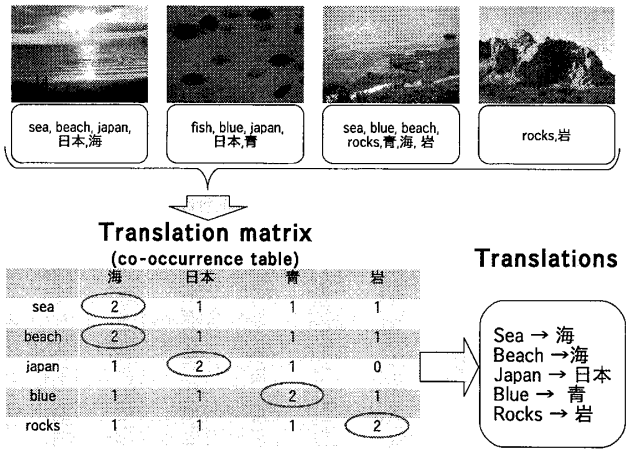

図 2　大規模言語及び小規模言語間での共起頻度を用いた変換マトリクス

画像とタグを収集し，大規模言語及び小規模言語を分類する．タグ付与プロセスでは，大規模言語によって構成される学習データを用いてタグ付与モデル (大規模言語モデル) を構築する．さらに，大規模言語モデルによって得られたタグ付与結果を小規模言語に変換することで小規模言語のタグを付与する．

4.1　学習データ収集プロセス

学習データを収集するために，WEB 上の画像共有サイトを利用して学習データを収集する．本プロセスでは，事前に学習データを収集するための検索クエリを設定し，検索 API を利用して画像及びタグを収集する．

検索クエリによって得られた検索結果を大規模言語及び小規模言語に分類するために，画像に付与されたタグに対してフィルタリングを適用する．フィルタリングによってタグを分類し，各言語タグが付与された画像を両言語データベースに関連付けて登録する．

尚，本プロセスでは画像共有サイトである Flickr[*1] を利用し，API[*2] により画像及びタグを収集する．

4.2　タグ付与プロセス

本プロセスでは文献 [3] に従い，大規模言語による学習データを用いて大規模言語モデルを構築し，入力画像に対して大規模言語タグを付与する．得られた大規模言語タグを小規模言語タグに変換するために，辞書に基づく変換方法と，学習データ内の共起頻度に基づく変換方法の 2 種類の方式を用いる．

尚，大規模言語モデルを構築する際に画像から SIFT 特徴量 [6] を抽出し，*BoVW*(Bag-of-Visual-Words)[7] を画像特徴量として用いる．さらに，文献 [8] に記載されている色相 (color moment)，エッジ方向ヒストグラム (edge direction histgram)，LBP(local binary pattern) を特徴量としたヒストグラムを画像特徴量として利用する．

4.3　小規模言語への変換方法

本節では，大規模言語を小規模言語に変換するための方法として，辞書に基づく変換方法及び，共起頻度に基づく変換方法について説明する．

大規模言語タグに対する辞書に基づく変換方法では，大規模言語を小規模言語に変換するための変換辞書を用意する．大規模言語モデルによって得られた大規模言語タグを変換辞書に入力することで，小規模言語におけるタグを得る．尚，本システムでは，変換辞書として，google より提供されている API[*3] を利用して大規模言語モデルによって得られた確率値の上位から順次小規模言語タグに変換する．尚，変換結果において既に同一の小規模言語タグが存在する場合には，変換結果として利用せずに大規模言語モデルの順位を繰り上げて変換を適用し，小規模言語タグを出力する．

次に，学習データ内での共起頻度に基づく変換方法について説明する．まず，大規模言語 DB 及び小規模言語 DB より，大規模言語と小規模言語によるタグが同じ画像で共起している画像を抽出する．抽出した画像群に付与されているタグの共起頻度を計算することで変換マトリクスを作成する．図 2 に大規模言語及び小規模言語間での変換マトリクスの概要図を示す．

大規模言語 L_A 及び小規模言語 L_B にて使用されているタグ集合 W_A，W_B を用いて変換マトリクス $T_{L_A \to LB}$ を表現すると式 1 のように表せる．

$$T_{L_A \to L_B} = \begin{bmatrix} t_{(1,1)} & \cdots & t_{(1,l)} & \cdots & t_{(1,L)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{(k,1)} & \cdots & t_{(k,l)} & \cdots & t_{(k,L)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{(K,1)} & \cdots & t_{(K,l)} & \cdots & t_{(K,L)} \end{bmatrix} \quad (1)$$

ここで、要素 $t_{(k,l)}$ は，タグ w_{kA} と w_{lB} の共起頻度を表している．

最終的に，大規模言語によるタグ結果を確率値の高いタグから順次変換マトリクスを参照し，最大となる共起頻度を示す小規模言語へと変換する．尚，既に同一の小規模言語タグが付与されている場合には，共起頻度を繰り上げて変換する．

[*1] http://www.flickr.com
[*2] http://www.flickr.com/services/api/
[*3] http://translate.google.com

表 1　学習データ収集用検索クエリ

英語	日本語
car	車
dog	犬
fireworks	花火
1 flower	花
food	食べ物, 食物
hanami	花見
ski,skiing	スキー，スキー場
sumo	相撲
tokyotower	東京タワー
sea	海
bird	鳥
bike	自転車

5　評価実験

本章では，学習データとして収集した画像群における大規模言語と小規模言語のタグ情報量の違いを検証する．さらに，提案システムの有効性を主観評価実験により検証する．

5.1　タグ情報量比較実験

5.1.1　画像データ収集・タグ処理

本実験の対象言語として英語及び日本語を採用し，事前に設定した検索クエリ (表 1) を用いて画像及びタグを収集した．収集したタグを対象に，英語と日本語のタグを分類し，両言語のタグ情報量の違いを比較した．ここで，英語検索クエリによる検索結果が膨大となったため，収集時間の短縮のためにすべての英語検索クエリと"Japan"及び"Japanese"を組み合わせて画像を収集した．収集対象となる期間は 2005 年 10 月 1 日から 2009 年 10 月 1 日とした．

また，収集した画像に付与されているタグから，ノイズとなるタグを削除するためにフィルタリング処理を適用した．具体的には，ストップワードを削除し，日本語と英語で常用されていない文字を含むタグを削除した．また，フィルタリングの結果，全てのタグが削除された画像は分析の対象外とした．

5.1.2　収集結果・分析

英語及び日本語検索クエリによって収集したタグに対して，フィルタリングを適用した最終的なデータにおけるタグ総数，ユニークなタグ総数及び画像総数を表 2 に示す．表 2 からも明らかな通り，タグ総数及びユニークタグ総数において，英語と日本語のタグの合計は大きな差があることが分かる．英語及び日本語の全体における画像一枚あたりの平均タグ数は，それぞれ 8.9，3.8(tag/image) となっていた．また，画像投稿者数を調査したところ，英語タグが付与された画像の投稿者数は 16,563 人，日本語を対象にした場合は 5,403 人となっていた．一人当たりが投稿する画像数は英語が 17.6 枚，日本語が 27.0 枚となり，日本語が多いことが判明した．

この様に，少ない人数によって付与された日本語タグに比べて，英語においては多くの投稿者によってタグが付与されていることからタグ情報が多様であると言える．これは表 2 のユニークタグ数においても確認できる．さらに，一枚あたりの画像に付与されている英語タグの総数は，日本語タグの総数に対して 2 倍以上であることから，英語タグを学習データとした場合のタグ情報量は多いことが確認できた．文献 [3] などで利用されている pLSA では，画像特徴量とタグの共起確率に基づいてタグ付与モデルを構築しているため，自動タグ付与モデルを構築する場合に，英語の学習データを用いた自動タグ付与モデルの精度が，日本語を学習データとしたモデルよりも精度が高くなる事が予想できる．

表 2　検索結果におけるタグ総数，ユニークタグ総数及び画像総数

	タグ総数	ユニークタグ数	画像数
英語	2,849,658	46,302	291,628
日本語	558,867	16,793	145,763

5.2　主観評価実験

5.2.1　実験システム

本実験では，提案システムとして 4.3 節に記載した辞書変換及び共起変換を用いたシステムを構築した (辞書変換システム，共起変換システム)．さらに，比較用として，自動タグ付与モデルの学習データとして，日本語を用いたタグ付与システム (日本語システム) を構築し，被験者による主観評価に基づいて比較する．

実験データは，5.1.2 節の収集結果に基づいて，英語を大規模言語，日本語を小規模言語と設定し，両言語が同時に付与されている画像群を利用した．両言語が同時に付与されている画像は 116,273 枚存在しており，日本語が付与された画像群の 80% の画像において英語のタグが付与されていた．これら画像群を自動タグ付与モデル構築及び，変換マトリクス作成の学習データとして用いた．

PLSA の学習データとして用いるために全ての画像から，*BoVW* 特徴量 300 次元，色相特徴量 150 次元，LBP250 次元，エッジ方向ヒストグラム 73 次元の合計 773 次元の特徴量を抽出した．*BoVW* 特徴量は実験データからランダムに 10% の画像を選択し，SIFT 特徴量を抽出したのち，k-means 法によりコードブックを作成した．各言語において 90% の画像を教師データとし，10% のデータをテストデータとして利用した．

5.2.2　実験方法

テストデータは表 1 に記載の各項目から 5 画像をランダムに選択し，合計 60 画像を実験に用いた．全ての画像を実験用システムに入力し，各システムのタグ付与結果から得られたタグの確率値上位 10 位までのタグを実験に用いた．(合計 30 タグ)

12 名の被験者に対して全ての画像と付与されたタグを閲覧してもらい，全てのタグについて評価を付与した．評価基準は，correct，incorrect，unknown の三項目で，各項目の評価基準は以下の様に設定した．

- correct:タグが画像の内容を表現している
- incorrect:タグが画像の内容を表現していない
- unknown:画像からは判断できない

さらに，被験者は画像一枚毎に全てのシステムが付与したタグに対して，精度のよい順に順位を付与することで，ランキング評価を実施した．

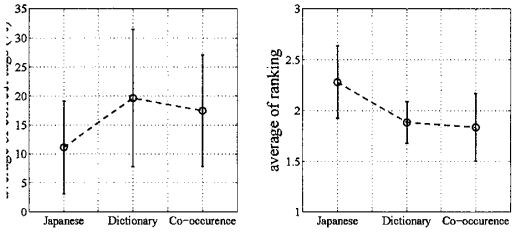

図 3　主観評価実験結果 (左：タグ評価結果平均及び標準偏差，右：ランキング評価結果平均及び標準偏差)

5.2.3　実験結果・考察

まず，システムによって付与されたタグに対する被験者の評価から，画像毎の correct 評価の割合の平均を計算した．さらに，全体に対する平均と標準偏差を図 3 に示す．日本語システムでは平均 11.1%(標準偏差：7.98)，辞書変換及び共起変換システムの平均はそれぞれ 19.8%,17.5%(標準偏差：11.8，9.58) となった．さらに各システムに対するランキング評価結果の平均は，日本語タグ付与システムが 2.3(標準偏差：0.36)，辞書変換システム及び共起変換システムでは，それぞれ 1.9，1.8(標準偏差：0.20，0.33) となった．また，t 検定により，日本語システムのタグ精度及びランキング評価結果に対し，辞書変換及び共起変換システムの評価結果の平均の差は統計的に有意であることを確認した．($\alpha = 0.05, p < 0.01$)

これら結果より，大規模言語による豊富な学習データを利用してタグ付与モデルを構築し，得られたタグを小規模言語に変換することで，小規模言語による少ない学習データを用いた自動タグ付与システムに比べてタグ付与精度が向上可能であることが確認でき，提案システムの有効性が確認できた．

また，辞書変換システムと共起変換システムの精度平均に統計的な差は確認されず，同等な主観評価精度であったと言える．両システムにおいて付与されたタグ結果の例を表 3 に示す。尚，”辞書”行及び”共起”行に記載されている例は，各変換結果が他方に対して高評価であった例を示している．共起変換システムでは，正確な対訳を付与する場合には向いていないため，写真に付与された英語のタグを日本語に変換することでよい評価が得られるケースでは評価が低い傾向があった．例えば，”street”に対する辞書変換結果は”ストリート”となったが，共起変換結果は，”東京”となっており，付与された英語タグが一般名詞であり，画像に移っている物体を表現する場合には辞書変換が優位である傾向が見られた．

しかし，共起変換では辞書に登録されていない意味や状況に変換できるメリットがある．例えば”cherry blossom”では，辞書には花見と言う意味は登録されていない．しかし，画像共有サイトでは主に花見時の写真がアップロードされており，写真が撮影された状況を考慮した変換結果となった．共起変換では，変換マトリクスの学習データとして実際に画像に付与された両言語のタグを用いており，日本語で利用されやすい類義タグに変換されることで主観評価が上がる傾向が見られた．今後，辞書変換と共起変換結果をうまく統合し，使い分けることでより高い精度が期待できる．

表 3　辞書変換及び共起変換高評価例

	英語タグ	辞書変換	共起変換
辞書	street	ストリート	東京
	bloom	満開	植物
	nature	自然	鳥
	英語タグ	辞書変換	共起変換
共起	cherry blossom	桜	花見
	tower	塔	東京タワー
	tournament	トーナメント	大相撲

6　まとめ

本稿では，学習データに混在する言語の中から小規模言語によるタグを付与するために，大規模言語によって構築された自動タグ付与モデルの結果を小規模言語へと変換する自動タグ付与システムを提案した．評価実験において，画像共有サイトから収集したタグ情報量が，英語と日本語で大きく異なることを確認した．さらに，主観評価実験によって日本語を学習データとしたタグ付与システムに比べ，英語を学習データとして，日本語に変換することで，評価値が向上されることを確認した．

また，大規模言語を小規模言語に変換する方法として，辞書に基づいて変換する方法と，学習データ内の両言語タグにおける共起頻度に基づく方法では，両者ともに主観評価値は同等の結果となった．辞書変換に比べて，共起変換を用いた場合では，辞書登録されていない単語への対応が可能である事が確認できた．

参考文献

[1] J. Li and J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal., 25(9), 1075-1088, 2003.

[2] F. Monay and D. Gatica-Perez. Modeling semantic aspects for crossmedia image indexing. Pattern Analysis and Machine Intelligence, IEEE Tran. on, 29(10):1802-1817, Oct. 2007.

[3] F. Monay and D. Gatica-Perez, PLSA-based image autoannotation: constraining the latent space. In ACM Multimedia, pp. 348-351. ACM, 2004.

[4] S. Romberg, E. Horster and R. Lienhart,:Multimodal plsa on visual features and tags, In Proc. of ICME 2009, pp. 414-417, 2009.

[5] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1):177-196, 2001.

[6] D. G. Lowe, Object recognition from local scale-invariant features. In ICCV, pp. 1150-1157, 1999.

[7] G. Csurka, C. Bray, C. Dance and L. Fan, Visual categorization with bags of keypoints. in Proc. of ECCV Workshop on Statistical Learning Computer Vision, pp. 59-74, 2004.

[8] A. Yanagawa, W. Hsu and S.-F. Chang. Brief descriptions of visual features for baseline trecvid concept detectors. Number Columbia University ADVENT Technical Report 219-2006-5, July 2006.