

# 画像・文章間の類似度学習による画像説明文の自動生成

牛久 祥孝<sup>†</sup> 原田 達也<sup>††</sup> 國吉 康夫<sup>†</sup>

<sup>†</sup> 東京大学大学院情報理工学系研究科

〒 113-8656 東京都文京区本郷 7-3-1

<sup>††</sup> 東京大学大学院情報理工学系研究科 / JST さきがけ

〒 113-8656 東京都文京区本郷 7-3-1

E-mail: <sup>†</sup>{ushiku,harada,kuniyosh}@isi.imi.i.u-tokyo.ac.jp

あらまし 爆発的に増加するマルチメディアデータを背景として、データの「検索」や「理解」の為の技術が求められている。何れにしても、データの内容を幾つかのラベルで表現するのみならず、それらの関係を包含する自然言語の文として説明できる技術が有用である。だが、事物間関係や文法的な知識を獲得する難しさから、先行研究は殆ど存在しない。本研究では、画像とそれを説明する文章を学習データとして、入力画像と扱う内容が類似した画像に付随する文章に注目し、説明文を再構築することでこれらを解決する。画像と説明文からなるデータセットで提案手法による説明文の生成実験を行った結果、提案手法が画像・文章から画像説明文を生成しうることを示した。

キーワード 画像説明文生成, 一般物体認識, 自然言語処理, 確率的正準相関分析, マルチスタックビームサーチ

## 1. はじめに

近年の情報技術の発展により爆発的に増加している画像や動画などのマルチメディアデータを効率的に活用する為に、個々のデータをユーザが「検索」し「理解」できる為の技術が求められている。何れの場合でも、マルチメディアデータの内容を幾つかのラベルで表現するのみならず、それらの関係を包含する自然言語の文として説明できる技術が有用である。

「検索」の技術は文書データを対象としたものをきっかけとして盛んに取り組まれている。対象がマルチメディアデータである場合、入力するクエリとしては大きく2つの選択肢がある。ひとつは同じ様態のマルチメディアデータである。たとえば画像検索なら、例となる画像をクエリとして類似した画像を検索する。あるいは音声検索なら、やはり例となる音声や音楽のデータをクエリとして所望の音声を検索する。このようにサンプルをクエリとするアプローチに対して、もう1つの選択肢は自然言語をクエリとするものである。現在の商用検索エンジンでは、単語によるクエリを受け付けてマルチメディアデータを検索している。

「理解」の技術としては要約技術があるが、これも自然言語データを対象としたものが端緒である。マルチメディアデータを対象とした場合には、同じ様態のデータで出力するか、自然言語として出力するかの選択肢がある。例えば動画を入力すると、前者では主要な場面のみからなる短い動画が出力され、後者の場合は動画の内容を説明する単語や文が出力される。

自然言語によるマルチメディアデータの「理解」では、独立な複数のラベルによる出力よりも、それぞれの関係

性を含んだ説明文としての出力が有用である。また、マルチメディアデータを自然言語化して検索可能にする場合でも、データの文としての表現によって、従来のばらばらな単語に加えて其々の関係性を含めた柔軟なクエリでの検索が実現できる。つまり、マルチメディアデータの「理解」と「検索」には文の体を為す出力が有用である。

マルチメディアデータを文で説明する研究は音声要約で多く行われている[1]が、これは音声認識によるもので、発話以外の音を含むデータを要約するものではない。そして動画要約では動画の再編集が殆どであり、自然言語出力を行う場合は字幕の文字認識や会話の音声認識に依存している。これでは動画像の内容を直接述べているとは言い難いが、一般的な動画・画像から説明文を生成する先行研究は殆ど無い。

画像においても単語間の関係性は存在する。これらを文の形態で表わせれば、今までにないデータ理解と所望のデータのより容易な検索が実現できる。画像を説明するためには、画像に含まれる事物の推定だけでなく、それらの関係の推定と正しい文法の運用が求められる。たとえば数万種の事物の一对一関係だけでも数億の事物間関係が存在することになり、これらの推定は非常に困難である。従って数少ない先行研究[2]では、これらの画像と文章にいくつかのラベルを付与し、入力画像のラベル推定を経て相応しい文章を検索している。一般的な画像に説明文を出力するにはより多くの学習データが必要であり、近年では Web の膨大なデータを用いる画像認識の研究も盛んである[3]。しかし Web では画像とキャプションから画像・説明文ペアを収集できるが、それらへのラベルは人手で付与せざるを得ず、高いコストが発生する。

本研究では、そのような画像・文章からなるデータを

用いて一般的な画像の説明文を生成する手法を開発する。画像の説明文を生成するために、入力画像と扱う内容が類似した画像に着目する。これらに付随する文章には、入力画像の一部を説明するような表現が存在し、そこには画像に含まれる事物だけではなく、それらの間の関係と文法的な知識が既に埋め込まれている。このような表現を複数の文から抽出し、文を再構成すれば、画像を内容的にも文法的にも正しく説明できる筈である。そこで本研究では、入力画像と類似した画像の訓練サンプル文を再構成するアプローチを提案する。また、説明文の評価として適切な評価指標の検討を行う。

本研究のコントリビューションを纏めると、次のようになる。1) 画像と文章のみからなるデータセットによる画像説明文生成を実現する新規アプローチの提案と、提案手法への定量的な評価。2) 画像の説明文生成に対して人手による評価と自動評価指標の相関の検討。

## 2. 関連研究

動画要約の枠組みにおいて自然言語ベースの出力を行うものは、スポーツ動画のみを対象とする等のドメイン限定や、音声認識によって抽出した発話・文字認識によって抽出した字幕のみを用いる [4] 等のモダル限定によって事物や事物間関係の推定や文法知識の獲得といった問題を回避している。ただし、一般的なマルチメディアデータの内容を説明する自然言語は出力できない。これは画像の説明文生成についても同様である。Yao ら [5] は、監視カメラに映る画像やドライブ中の車上から撮影した画像の説明文出力を行っている。セマンティック Web の枠組みで用いられる OWL (Web Ontology Language) に入力画像を変換し、OWL で記述された画像の内容を自然言語に変換する。ただし、出力された文章の例は 6 例にとどまり、定量的な評価は全くなされていない。

一般的な画像の説明文を出力するごく僅かな先行研究 [2] では、画像と文章がペアになっているデータセット PASCAL Sentence を提供し、この上で画像説明文生成を行っている。PASCAL Sentence は画像認識ワークショップの PASCAL で配布されたデータセットの一部、20 カテゴリの画像を 50 枚ずつ抽出し、各画像に対してクラウドソーシングの Amazon mechanical turk を利用して 5 文前後の文を付与している。各文は平均で 12 単語弱の長さであり、画像に写っている内容を直接記述している。

[2] では、訓練サンプルの画像と文章に対してトリプレット状のラベルを付与している。それぞれ「主題」「動作」「シーン」を記述するラベルであり、このラベルと画像もしくはラベルと文章の関係を学習している。入力画像から一旦トリプレットを推定し、類似するトリプレットを持つ文章を出力している。だが、この様に画像や文章に対して明示的な意味を与える作業は大きな労力を伴う。また、画像に「主題」「行動」「シーン」のトリプレットだけで表現できない内容があったとしても、それを説



図 1 提案手法の概要。

明する文は生成できない。

加えて、画像への説明文生成が研究領域として発展する為には、関連手法との客観的な性能比較を容易に行える状況が必要不可欠である。[2] でもやはり客観的な評価の必要性に触れられており、トリプレット推定精度の自動評価と最終的な出力文に対する人手による評価が行われている。ただ、出力文に対する自動評価は行われておらず、他の手法との相対的な比較は困難である。

## 3. 訓練サンプル文再構成アプローチ

画像説明文を生成するためには第 1 章で述べた通り、画像に含まれる事物とそれぞれの関係を推定した上で、文法的に正しく単語を並べる必要がある。本研究では、画像の説明文を生成するにあたって、入力画像と扱う内容が類似した訓練サンプル画像に着目する。これらのサンプル画像の説明文には、入力画像の一部を説明するような表現が存在すると期待される。この表現には画像に含まれる事物の名前だけではなく、実物間の関係と文法的な知識が既に埋め込まれている。このような表現を複数の文から抽出し、文を再構成すれば、画像を内容的にも文法的にも正しく説明できる筈である。そこで本研究では、入力された画像に対して何らかの方法で類似した画像を検索し、付随する説明文を再構築することで入力画像を説明するアプローチを提案する。

この章では、提案するアプローチの具体的な手法について述べる。類似した訓練サンプルの検索としては、画像特徴量空間での近傍検索を考える。そこでまず、画像特徴量間の距離計量を考慮した特徴量空間の改善を 3.1 節で行う。3.2 節では、特徴量間距離と内容の類似度との間に依然として残るギャップを解消する為に説明文間の類似度を用いる手法について述べる。ここまでで定義された類似度に基づいて、入力画像と類似する訓練サンプルの検索を行う。獲得した訓練サンプルの説明文から入力画像を説明する表現を抽出する方法を 3.3 節で説明する。このような表現を組合せて説明文を生成する方法について、最後の 3.4 節で述べる。

### 3.1 特徴量の適切な距離計量の埋め込み

画像や文章から抽出された特徴量は、個々に適切な距離計量を持つと考えられる。予めそのような距離計量

微量空間へ埋め込み、各特徴量をユークリッド距離で計量できる様にする.[6]では、全訓練サンプルと一部の訓練サンプルからなるグラム行列を用いたカーネル主成分分析 (KPCA: Kernel Principal Component Analysis) を用い、特徴量間の非線形な距離計量が埋め込まれたユークリッド空間を獲得している。全  $N$  ペアの訓練サンプル間でのグラム行列を用いるとなると、計算量が  $O(N^2)$  になってしまう。ここでは訓練サンプルの一方を  $n < N$  だけに制限してスケーラビリティと精度のバランスを保ちながら、特徴量空間に非線形な距離計量を埋め込む。

特徴量  $x_i$  と  $x_j$  のカーネル関数を  $K(x_i, x_j)$  と表記する。ある特徴量  $x$  と  $n$  個のサンプルとのカーネル関数の出力を並べたベクトルを  $k_x = (K(x, x_1), \dots, K(x, x_n))$  として定義すると、サブセット上の KPCA は全  $N$  サンプルについて  $k_1, \dots, k_N$  を計算し、通常の PCA を行う手続きと等価になる。また、 $K(x_i, x_j)$  として [6] では L1 距離、L2 距離、 $\chi^2$  距離などの距離カーネルを用いている。これらの距離を  $dist(x_i, x_j)$  と表わす時、 $K(x_i, x_j) = \exp(-\frac{1}{2d}dist(x_i, x_j))$  と定義されたカーネル関数を用いる。ここで、 $\bar{d}$  は距離の平均である。

本研究では、距離に基づくカーネルとして L1 カーネル、L2 カーネル、 $\chi^2$  カーネル、Bhattacharyya カーネル、Jeffrey カーネル、Hellinger カーネル、ヒストグラムインターセクションカーネルを、構造データである文章のカーネルとしてストリングカーネルを検討した。実験的にいずれかの特徴量と組合せたときに訓練サンプル検索精度が高かったものについてのみ説明すれば、L1/L2 カーネルはそれぞれ L1 距離または L2 距離に基づく距離カーネルである。Bhattacharyya カーネルは Bhattacharyya Divergence に基づく距離カーネルで、離散確率分布間の場合は  $dist(x_i, x_j) = -\log \sum_k \sqrt{x_{ki}x_{kj}}$  で距離が求められる。Jeffrey カーネルは Jeffrey Divergence に基づく距離カーネルで、この距離計量は Kullback-Leibler Divergence (KLD) が対称性を持つように拡張したものである。離散確率分布の場合は  $dist(x_i, x_j) = \sum_k x_{ki} \log \frac{2x_{ki}}{x_{kj}+x_{ki}} + x_{kj} \log \frac{2x_{kj}}{x_{ki}+x_{kj}}$  で求められる。ヒストグラムインターセクションカーネルは、ヒストグラム間類似度としてそれぞれの共通部分を計算する。ヒストグラムの各ビンでの共通部分が  $\min(x_{ki}, x_{kj})$  で求められることに留意すれば、 $K(x_i, x_j) = \sum_k \min(x_{ki}, x_{kj})$  で求められる。

### 3.2 訓練サンプルの検索

本研究の訓練サンプルは画像と文章のペアを為しており、類似サンプル検索のナイーブな手法としては、入力画像の特徴量と距離が近い特徴量をもつ画像の検索も考えられる。しかし、画像の見た目とその内容の間にはセマンティックギャップが存在し、画像特徴量間の距離は必ずしも画像間の内容の類似度と相関をもたない。

そこで、説明文間の類似度を利用した画像間類似度の

改善を考える。Ushiku ら [7] は、入力画像から画像と文章のペアを検索する為の類似度学習手法を提案している。[7]では、Web の一般的な画像データのようにに文章が付随するようなデータセット上で類似画像検索を行う際は、画像の特徴量間距離のみでの近傍検索よりも、文章の類似度を反映した空間での近傍検索がより高精度だと報告されている。この手法は後述するようにスケーラビリティにも優れている。

前節で距離が埋め込まれた画像/文章特徴量を改めて  $x$  や  $y$  とし、潜在変数  $z$  で表現される潜在空間を仮定する。ある内容に基づく  $z$  の画像や文章表現として  $x$  ないし  $y$  が確率的に出現していると捉える。すると逆に、画像・文章ペアの訓練サンプルについては確率  $p(z|x, y)$  で、入力画像については確率  $p(z|x)$  で  $z$  を推定できる。よって、目的の類似画像検索はこれらの確率分布の確率間距離に基づく近傍検索となる。これらの確率分布を以下の正規分布で表わす。

$$z|x, y \sim \mathcal{N}(\mu_{xy}, \Phi_{xy}), z|x \sim \mathcal{N}(\mu_x, \Phi_x). \quad (1)$$

正規分布など指数型分布族では確率間擬距離として KLD を用いる。入力された画像から推定される潜在変数と訓練サンプルの画像・文章ペアから推定される潜在変数が、それぞれ平均  $\mu_x$  分散  $\Phi_x$  ないし平均  $\mu_{xy}$  分散  $\Phi_{xy}$  の独立同一正規分布であると仮定する。このとき、入力画像  $x_q$  から訓練サンプル  $\{x_t, y_t\}$  への KLD から定数項を除き、CCD を次のように定義する。

$$CCD(x_t, y_t|x_q) = \left\| \Phi_{xy}^{-\frac{1}{2}}(\mu_x - \mu_{xy}) \right\|^2. \quad (2)$$

なおこれは、訓練サンプルの潜在変数を  $r_t$ 、文章の付随しない画像である時の潜在変数を  $r_q$  としたとき、

$$r_t = \Phi_{xy}^{-\frac{1}{2}}\mu_{xy}, r_q = \Phi_x^{-\frac{1}{2}}\mu_x, \quad (3)$$

と定義された  $r$  でのユークリッド距離に基づく近傍検索と等価である。

以上から、 $p(z|x, y)$  と  $p(z|x)$  の平均  $\mu_{xy}$ 、 $\mu_x$  と分散  $\Phi_{xy}$  を推定する問題へと帰着された。ここで、 $p(z|x, y)$  と  $p(z|x)$  を正規分布としたため、 $p(z)$  と  $p(x|z)$ 、 $p(y|z)$  も正規分布で表現できる。これらを、 $z \sim \mathcal{N}(0, I)$ 、 $x|z \sim \mathcal{N}(W_x z + m_x, \Psi_x)$ 、 $y|z \sim \mathcal{N}(W_y z + m_y, \Psi_y)$  として、画像と文章がペアになっているデータを用いて最尤推定を行う。これは、確率的正準相関分析 (pCCA: Probabilistic Canonical Correlation Analysis) [8] で行う最尤推定と全く同じものである。正準相関分析 (CCA: Canonical Correlation Analysis) は、画像特徴量ベクトル  $x$  と文章特徴量ベクトル  $y$  に対し、正準変量  $s = A^T(x - \bar{x})$  と  $t = B^T(y - \bar{y})$  の相関を最大にする行列  $A$  と  $B$  を求めるものである。pCCA は CCA を確率的に解釈したもので、それぞれの平均と分散は、

$$\mu_x = M_x^T A^T (x - \bar{x}), \Phi_x = I - M_x M_x^T, \quad (4)$$

$$\mu_{xy} =$$

$$\begin{pmatrix} M_x \\ M_y \end{pmatrix}^T \begin{pmatrix} (I - \Lambda^2)^{-1} & -(I - \Lambda^2)^{-1} \Lambda \\ -(I - \Lambda^2)^{-1} \Lambda & (I - \Lambda^2)^{-1} \end{pmatrix} \begin{pmatrix} A^T (x - \bar{x}) \\ B^T (y - \bar{y}) \end{pmatrix}, \quad (5)$$

$$\Phi_{xy} = I -$$

$$\begin{pmatrix} M_x \\ M_y \end{pmatrix}^T \begin{pmatrix} (I - \Lambda^2)^{-1} & -(I - \Lambda^2)^{-1} \Lambda \\ -(I - \Lambda^2)^{-1} \Lambda & (I - \Lambda^2)^{-1} \end{pmatrix} \begin{pmatrix} M_x \\ M_y \end{pmatrix}, \quad (6)$$

と解析的に求められる。なお、 $\bar{x}$  と  $\bar{y}$  は学習に用いた特徴量の平均である。また、 $M_x$  と  $M_y$  は  $M_x M_y^T = \Lambda$  かつスペクトルノルムが 1 未満となるような任意の行列である。ここでは、CCD が用いている  $0 < \beta < 1$  なる  $\beta$  で  $M_x = \Lambda^\beta$ ,  $M_y = \Lambda^{1-\beta}$  と定義した。また、 $A$  と  $B$  は通常の CCA と同じ一般固有値問題  $R_{XY} R_Y^{-1} R_{YX} A = R_X A \Lambda^2$  や  $R_{YX} R_X^{-1} R_{XY} B = R_Y B \Lambda^2$  の解である。ただし、 $\Lambda^2$  は固有値を対角要素とする対角行列であり、ただし、 $\Lambda^2$  は固有値を対角要素とする対角行列であり、 $R_X = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$ ,  $R_Y = \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T$ ,  $R_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})^T = R_{YX}^T$  で求められる。CCA によって得られる空間は正準空間と呼ばれ、両変量の分布を考慮した潜在空間となっている。

このように CCA は特徴量次元と同じ次元数の一般化固有値問題を解けばよいので、データ数に対するスケラビリティに優れている。pCCA は CCA によって得られる正準相関を用いて、2 つの正準空間を上手く融合させる手法とも捉えられる。確率的に潜在変数を解析する枠組みとして、確率的潜在意味分析や潜在的ディリクレ配分などもあるが、反復計算による最適化が不要で、大域的最適解が求まる点では pCCA の方が優れている。

### 3.3 訓練サンプル文中のフレーズ選択

前節で述べた CCD による近傍検索で得た訓練サンプルの説明文には、入力画像自身の事物やそれらの関係を文法的に正しく説明するフレーズ（連続単語列）の存在が期待できる。つまり、入力画像を説明する確率の高いフレーズの組み合わせによって、文法的にも内容的にも正しい画像説明文が生成できると考える。近傍検索によって得た訓練サンプルに含まれる文の集合を  $\mathcal{W}$  とし、それぞれの文を  $w = w_1^l = (w_1 w_2 \dots w_l)$  と表わす。  $l$  は文の単語数で、 $w_i$  は文頭から  $i$  番目の単語である。入力画像の特徴量  $x_q$  から、ある訓練サンプル画像  $x_i$  と説明文  $y_i$  のペア  $s_i$  が検索される確率を  $p(s_i|x_q)$ 、ペア  $s_i$  から、単語が一つ以上並んだあるフレーズ  $p_{hr}$  が選択される確率を  $p(p_{hr}|s_i)$  とすると、入力画像  $x_q$  に対して各フレーズ  $p_{hr}$  を選択する確率は以下のように求められる。

$$p(p_{hr}|x_q) = \left( \sum_i p(p_{hr}|s_i) p(s_i|x_q) \right). \quad (7)$$

入力画像の特徴量  $x_q$  から、ある訓練サンプル画像  $x_i$

と説明文  $y_i$  のペア  $s_i$  が検索される確率  $p(s_i|x_q)$  を、ソフトマックス関数を用いて、

$$p(s_i|x_q) = \frac{\exp(-\text{CCD}(s_i|x_q))}{\sum_i \exp(-\text{CCD}(s_i|x_q))}, \quad (8)$$

と定義する。次に、それぞれのペア  $s_i$  から、単語が一つ以上並んだあるフレーズ  $p_{hr}$  が選択される確率  $p(p_{hr}|s_i)$  を tfidf 重み付け [9] を用いて、

$$p(p_{hr}|s_i) = \frac{\text{tfidf}(p_{hr})}{\sum_{p_{hr} \in \mathcal{W}_i} \text{tfidf}(p_{hr})}, \quad (9)$$

と定義する。 $\mathcal{W}_i$  は訓練サンプル画像と説明文のペア  $s_i$  が持っている文の集合である。tfidf 重み付けは、その文におけるフレーズの頻度 (term frequency) に、それぞれのフレーズが全文書で出現する頻度 (document frequency) の逆数 (inverse) をかけたものである。冠詞や be 動詞のように、どの文にも頻繁に出現するような単語やフレーズの重みを下げ、それぞれの文に特徴的なフレーズに大きなスコアを与える指標である。

このように、あるフレーズが入力画像を説明する確率を、各訓練サンプル画像と説明文のペア  $s_i$  を仮説としたベイズ最適な枠組みで求める。確率の高いフレーズを組合せて説明文を構築する方法については、次節で述べる。

### 3.4 フレーズの組合せによる文生成

説明文を出力するために、前節で求められる選択確率の高いフレーズを組み合わせる。確率の高いフレーズを出発点として、そのフレーズの左右にやはり高確率のフレーズが生じるような単語を並べていくアプローチを考える。これは、統計的機械翻訳のデコーディングに相当する。ある言語の文章  $f$  を目標の言語の文章  $e$  に翻訳する過程は、対数線形モデル [10] では次の式の  $\arg \max$  を求める処理（デコーディング）とみなせる。

$$\hat{e} = \arg \max_e \sum_k \lambda_k \log f_k(e, f). \quad (10)$$

$\phi_k = \log f_k(e, f)$  は素性関数と呼ばれ、片方の言語もしくは両方の言語の情報を使用して算出する。特に  $f(f|e)$  は翻訳モデルと呼ばれ、対訳コーパスから獲得する。また、 $f(e)$  は言語モデルと呼ばれ、目標の言語で書かれた文書データセットから抽出した n-gram モデルなどがこれに該当する。

本研究では素性関数としてまず、前節で定義したフレーズ選択関数の対数を用いる。ただし、7 の定義そのものでは特定の支配的なフレーズが意味のない繰り返しを起こす可能性がある。たとえば、ある入力画像に対して "in a", "a place", "place in" が最もスコアの高い三つのフレーズとなった場合には "... in a place in a place in a place ..." という文が最も高いスコアを持ってしまう。そこで、あるフレーズ  $p_{hr}$  が単文で出現した回数の最大値を  $o_{max}(p_{hr})$ 、現在生成している文章で出現している

回数を  $o(p_{hr})$  としたとき,

$$\phi(p_{hr}|x_q) = \left(1 - \frac{o(p_{hr})}{o_{max}(p_{hr})}\right) \log \left( \sum_i p(p_{hr}|s_i) p(s_i|x_q) \right), \quad (11)$$

として既出フレーズの再利用を抑制する.

また, 基本的に文が伸長すると対数の和は低下するので, 2 つ目の素性関数として適切な長さの文を生成する為の文長スコアを導入する. これは同時に, 入力画像の内容を説明する文の長さをユーザがある程度指定できることを意味する. 所望の長さ  $l_{tar}$  の文を生成させるとき, 文長  $l$  の確率分布を  $l \sim \mathcal{N}(l_{tar}, \sigma)$  と仮定して, この対数を文長スコアとする.  $\sigma$  は定性的には文長の許容範囲を意味している. 本研究では,  $l_{tar} = 12$ ,  $\sigma = 1$  とする.

本研究は以上の 2 つの素性関数を用いる. n-gram による文法モデルなども使用できるが, まずは訓練サンプル文の再構成のみで生成された説明文の精度を検討する.

基本的に生成可能な文は語彙数の文長乗だけ存在する為, 全ての翻訳文候補を生成した上での比較は困難である. そこで, デコーディングにはグラフ探索の手法であるマルチスタックビームサーチが頻繁に採用される. 単語を左から並べる基本的な Left-to-right 展開でのグラフ探索は, ある位置  $i$  までの単語列  $w_1^i$  に基づいて次の単語  $w_{i+1}$  を選択する. 全体のスコアが最大となるパスを発見するグラフ探索法は種々あるが, ビームサーチは枝を次々と伸ばしていった時に上位  $s$  個のパスをスタックに保存しておく方法である. スタックのサイズ  $s$  を調整すればグラフ探索の精度と計算量のバランスを変更できる.

ただし, 確率の積算に基づくスコアは枝が延びる程小さくなる. 結果として短い文のスコアが高くなってしまいう為, 十分な長さまで文が伸びない可能性がある. そこで, マルチスタックビームサーチでは各文長毎にスタックを用意する. 統計的機械翻訳の枠組みでは, 翻訳された日本語の単語数に応じてスタックが分けられる. 日本語で翻訳が未了の部分が無くなれば探索終了であり, 最後のスタックで上位にある文がシステムの翻訳文となる.

本研究では入力文ではない為, 統計的機械翻訳のマルチスタックビームサーチそのものを行う場合には大きくわけて二つの問題がある.

- 画像対文章のため, 単語の位置対応が存在せず, Left-to-right 展開が行えない.

- 画像の「未翻訳」部分の判定は難しく, 文生成終了のタイミングが判らない.

そこで, 本研究では以下の改良を行う.

- 両側へ伸長する文生成プロセスの導入.
- 生成文の文長を基準としたマルチスタックの採用.
- スコア付き終了可能インデックスの記憶.

提案手法では, まず入力画像から検索された訓練サンプルの文集合  $\mathcal{W}$  にある長さ  $l_s$  のフレーズについてスコ

アを計算する. そして, これらを種として同じ長さ  $l_e$  だけ Left-to-right および Right-to-left 展開する. スタックは現在生成中の文そのものの長さを基準として分ける.

そして, それぞれのスタックにある候補文にはスコア付き終了可能インデックスを付与する. 予め訓練サンプルの文集合  $\mathcal{W}$  において, 各文の先頭と末尾に終了文字 (EOS: End of Sentence) を付加する. 文の伸長途中に EOS へと繋がるフレーズが  $\mathcal{W}$  に存在する場合, その時点で EOS に繋がった時のスコアの大小に関わらず, 現在のインデックスと繋がった時のスコアを記憶しておく. 展開が終了したら, 次の式に従って候補文中から EOS で挟まれた部分単語列候補を抽出する.

$$w = \arg \max_{w=(w_{i_s} w_{i_s+1} \dots w_{i_e-1} w_{i_e})} \sum_k \phi_k(w_{i_s}^{i_e}|x_q), \quad (12)$$

ここで  $i_e$  は Left-to-right 展開中に発見した EOS へ接続可能なインデックス,  $i_s$  は Right-to-left 展開中に発見した同様のインデックスである.

## 4. 実 験

本研究は実験用データセットとして PASCAL Sentence [2] を用いる. 900 ペア (カテゴリそれぞれで 45) の画像・文章データを訓練用サンプル, 残る 100 ペアの画像・文章データを評価用サンプルとする. 訓練サンプルの選択をランダムに 5 回行い, 結果の平均を用いる.

以降の節で述べるような指標の下で訓練サンプル検索のパラメータを最適化し, 提案手法による説明文生成を行う. まず自動評価の妥当性について検証し, その後提案手法が生成する説明文の精度について実験する.

### 4.1 特 徴 量

画像特徴量としては, 画像の形状情報とテクスチャ情報の利用が求められる. そこで, おもに形状情報を対象とした特徴量として SIFT [11], テクスチャ情報を対象とした特徴量として HLAC [12], Gist [13], LBP [14] を採用する. SIFT は, 近傍の複数のコードブックで局所特徴量を表現する Locality-constrained Linear Coding (LLC) [15] によって画像全体の特徴量を得る. これは, 従来よく用いられている Bag of Features 表現より画像分類性能がよいことが知られている. また, テクスチャ情報はグレースケールや RGB などの色空間によって扱う情報が異なると考えられる. 本研究では, 大域的特徴量それぞれについて [16] で用いられている色空間 (グレースケール, RGB, Transformed color, Opponent color, C-invariant color) で表現された画像から特徴抽出を行う.

文章特徴量としては, 各文にどのような単語やフレーズが出現したかという情報を利用する. 具体的には, 文全体に出現した単語/フレーズが 1 つの成分に対応するベクトル空間モデルを採用する. 文に出現しない単語/フレーズに対応する成分は 0 になる. 出現する単語/フ

レーズに対する重み付けとしては、フレーズスコア関数にも用いた tfidf 重み付け [9] を採用する。各文においてその単語/フレーズがいかほどの割合で出現するかの頻度 (term frequency) のみでは "is" や "in a" のような意味を持たない単語/フレーズへの重み付けが行われてしまう。単語に関してはそのような単語を stop word として排除する選択肢もあるが、フレーズに関してそのような枠組みは存在しない。そこで、文書頻度の逆数 (inverted document frequency) を乗ずることで、多くの文に出現する単語/フレーズの重みを下げられる。本研究では、単語に対する tfidf と、2単語～7単語の連続単語列からなるフレーズに対する tfidf の計7種類の特徴量を抽出する。

#### 4.2 評価指標

この節では、訓練サンプル検索の精度と最終的な説明文出力の評価指標について、それぞれ述べる。

画像説明文の評価については、データセットとして画像に予め与えられていた参照文とどの程度一致しているかを定量化する。そこで、統計的機械翻訳における評価指標である BLEU (BiLingual Evaluation Understudy) [17] と NIST [18] を使用する。BLEU は最もスタンダードな指標で、n-gram マッチ率を unigram から 4-gram まで積算したものである。文章が短いとマッチ率の積が不当に高くなるので、参照訳より短い文章については文長ペナルティが課される。

$$BLEU = \exp \left\{ \sum_{n=1}^N w_n \log(p_n) - \max \left( \frac{L_{ref}^*}{L_{sys}} - 1, 0 \right) \right\}, \quad (13)$$

ここで、 $L_{sys}$  は生成した文の単語数、 $L_{ref}^*$  は参照文の単語数である。通常は複数の参照文が存在するので、最も生成された文の単語数と近いものを使用する。また、 $w_n$  は各 n-gram ごとの重みづけであり、通常は  $\frac{1}{N}$  である。 $p_n$  は生成された文中の n-gram で参照文中にも登場するものの割合である。NIST はこれを改良した指標で、マッチした n-gram それぞれの情報量の総和を用いる。

$$NIST = \sum_{n=1}^N \left\{ \frac{\sum_{all \ w_1 \dots w_n \ in \ sys \ output} \text{info}(w_1 \dots w_n)}{\sum_{all \ w_1 \dots w_n \ in \ sys \ output} (1)} \right\} \exp \left\{ \beta \log^2 \left[ \min \left( \frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}, \quad (14)$$

ここで、 $\beta = \log(0.5)/\log(1.5)^2$  である。info( $w_1 \dots w_n$ ) は  $w_1 \dots w_n$  のもつ情報量で、参照文全体での  $w_1 \dots w_{n-1}$  の生起回数を  $w_1 \dots w_n$  の生起回数で割ったものである。

また、提案手法では画像の説明文生成のために有用な表現を含む訓練サンプルを検索している。この検索部分のパラメータを最適化する為の指標としても BLEU と NIST を用いる。具体的には、入力画像に本来付随していた説明文と、検索された説明文との n-gram 一致度をそれぞれの定義で計算する。また、検索精度評価として良く用いられる MAP (Mean Average Precision) も検討する。これは、

表 1 訓練サンプル検索の評価。左:画像特徴量のみでの検索。右:文章特徴量と組合わせた CCD での検索。

| 特徴量  | 画像特徴量のみ |      |        | CCD  |      |       |
|------|---------|------|--------|------|------|-------|
|      | MAP     | NIST | BLEU   | MAP  | NIST | BLEU  |
| SIFT | 0.11    | 0.45 | 0.0066 | 0.14 | 0.49 | 0.014 |
| HLAC | 0.080   | 0.45 | 0.0064 | 0.13 | 0.47 | 0.012 |
| Gist | 0.081   | 0.45 | 0.0065 | 0.14 | 0.47 | 0.013 |
| LBP  | 0.084   | 0.45 | 0.0068 | 0.13 | 0.49 | 0.012 |

入力画像のカテゴリ情報と検索されたサンプルのカテゴリ情報の一致率を図るものである。次の式でカテゴリごとに求められる平均適合率  $AP = \frac{1}{N_c} \sum_{r=1}^k \text{Pr}(r) \text{Rel}(r)$  の平均値が MAP である。ここで、 $k$  は検索するデータ点数で本研究は前実験を通じて 5 である。 $N_c$  はカテゴリ  $c$  のデータ点数、 $\text{Pr}(r)$  は  $r$  番目までの検索結果のうち正しいカテゴリ  $c$  に属するデータの割合 (Precision)、 $\text{Rel}(r)$  は  $r$  番目の検索結果が正しいカテゴリ属する場合のみ 1 をとり、そのほかの場合には 0 となる関数である。ただし、この評価指標は画像・説明文データにカテゴリ情報 (つまり、ラベル) が付随していなければ計算できない点に注意されたい。本研究が目指す画像・文章のみからの説明文出力では、ラベルが必要な MAP 以外の指標でパラメータを最適化できることが望ましい。

#### 4.3 訓練サンプル検索の最適化

各画像特徴について、画像特徴量のみでもっとも評価指標が高くなるパラメータと、CCD に用いたときにもっとも評価指標が高くなるパラメータを探索した。CCD の成績は 3.2 節で述べたパラメータ  $\beta$ 、画像・文章量特徴量の主成分分析における打ち切り次元、pCCA の打ち切り次元の 4 パラメータに左右される。具体的には、 $\beta$  については 0.1 から 0.2 刻みで 0.9 まで、pCCA の打ち切り次元は画像・文章特徴量で次元数が少ない方に対し 2 割刻みで 2 割～10 割、画像・文章それぞれの特徴量の打ち切り次元については 20 次元刻みで最適なパラメータを探索した。画像特徴量の色空間や部分カーネル主成分分析のカーネル関数は任意の組み合わせを試した。

画像特徴量のみで訓練サンプルを検索した場合、文章特徴と統合して CCD で訓練サンプルを検索した場合の各評価指標の値を表 1 に示す。この表が示すとおり、各特徴量を単独で用いるよりも、CCD による検索のほうが精度よく訓練サンプルを検索できている。以降では、それぞれの特徴量で最も指標が高くなるパラメータを用いる。画像特徴量の場合は 4 種の特徴量を使用しているが、それぞれの最適なパラメータに基づく特徴量を単純になぎ合わせて用いる。

#### 4.4 生成された文の評価

前節で最適化した CCD パラメータと特徴量で説明文を生成する。提案手法で生成された説明文の一部を図 2 に示す。まずは自動評価と人手での評価の相関を示し、以

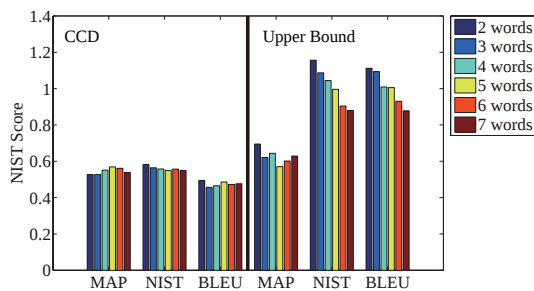


図3 Comparison with NIST score.

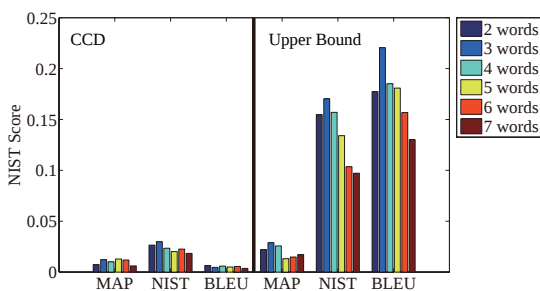


図4 Comparison with BLEU score.

降の自動評価のみでの比較の正当性を示す。その後、生成された説明文の例と種々の実験から考察を行う。

#### 4.4.1 生成された文の自動評価と人手による評価

本研究では訓練サンプル検索のための CCD パラメータを MAP, NIST, BLEU の 3 種の評価値それぞれで独立に最適化している。だが理想的には、それぞれの評価値で最良の距離が存在する。MAP では同カテゴリに属する訓練サンプルとの距離が 0 であり、その他が無限遠にある場合が、NIST/BLEU ではそれぞれの値が最も高い訓練サンプル 5 データとの距離が 0 であり、その他が無限遠にある場合が理想である。

そこで、3 種の評価値それぞれでパラメータ探索を行ってきた CCD によって近傍検索された訓練サンプルを使用した場合に加えて、それぞれ上記の距離による理想的な近傍検索を行えたとして得られる訓練サンプルを使用した場合でも説明文生成を行った。これらの理想的な距離計量に基づいた説明文生成性能は、訓練サンプル検索の精度によって説明文生成の性能が変動する範囲の上限値とみなせる。各画像について長さ 2 のフレーズを用いて計 6 通りの文を生成した。これらの文の順序をランダムに入れ替えた上で画像に付与したものを、10 人の被験者に対して 20 枚ずつ提示した。それぞれの画像について 2 名の被験者が評価し、計 100 枚の画像についてのべ 200 の自動評価と人手の評価を獲得した。人手の評価は、各説明文の適切さと流暢さをもとにした 5 段階評価とした。

自動評価値に関しては 6 種のシステムが 100 枚の画像に対して出力した説明文の評価全体で標準化を行い、人手による評価値に関しては被験者ごとの評価で標準化を行った。結果、評価間の相関値は人手による評価値と BLEU で 0.63, 人手による評価値と NIST で 0.82 であった。特に NIST で人手による評価値と強い相関がみられ、

BLEU と人手の評価にもかなりの相関がみられた。このことから、これらの自動評価に基づいた説明文生成の考察は妥当といえる。

#### 4.4.2 フレーズ長及び検索精度の説明文生成への影響

4.4.1 節で説明した 6 種類のシステムそれぞれについて、4.4.1 節では長さ 2 に限定していた使用フレーズ長を 7 まで増やしながら性能の変化を調べた。それぞれで生成された説明文に対する NIST のスコアを図 4 に、BLEU のスコアを図 5 に示す。それぞれ、左から CCD のパラメータを MAP, NIST, BLEU によって探索したもの、理想的な距離を MAP, NIST, BLEU に基づいて定義したものによって検索された訓練サンプルから、使用フレーズ長を 2~7 で変化させたときのスコアとなっている。なお、CCD で検索した訓練サンプルで長さ 2 のフレーズを用いて説明文を生成した場合、人手による評価が 3 を超えた説明文は MAP で 29 文/200 文、NIST で 30 文/200 文、BLEU で 9 文/200 文であった。意味のある説明文を生成できた割合が最大で 15% であったことになり、改善の余地は大きい。

改善の方針としてまず、訓練サンプル検索精度の向上が考えられる。CCD で訓練サンプルを検索した場合、理想的な距離に比べて BLEU で  $\frac{1}{5}$ , NIST でも約半分と大きな差がある。これは、訓練サンプル検索の精度が上がれば、生成される説明文もより正しくなることを示している。CCD による訓練サンプル検索の精度を改善させるには、特徴量の改良やデータセット規模の拡張が考えられる。[7] でも、画像・文章データセットにおいて規模が増大すると検索精度が上昇すると報告されている。CCD は距離学習の計算量はデータ量に対して一定でスケラビリティに長け、データセット拡張も可能である。

改善のもう一つの方針としては、文生成時に用いる素性関数の改良が考えられる。今回は単一長さのフレーズのスコアと、文長スコアの 2 つだけを用いている。複数のフレーズを扱う、n-gram モデルによる文法スコアを導入する、などによって、説明文の内容の適切さと文法の流暢さが向上すると考えられる。

またこれらの結果から、画像説明文生成のための CCD のパラメータチューニングには NIST を用いた場合に生成された説明文の NIST, BLEU が最も高くなる結果が得られた。本研究では [2] のように画像と文章以外に人手でラベルを付与する作業のコストを問題点の 1 つとして指摘している。提案手法そのものはこのようなラベル情報を一切使用しないが、MAP の計算時にはそれぞれのデータのラベル情報が不可欠となる。NIST による CCD のパラメータチューニングがより良い選択肢であると示されたため、提案手法はラベル情報をパラメータチューニングを含めた全場面で使用しないものと言える。

## 5. おわりに

本研究では、増大する画像の効率的な理解・検索のた

## GOOD EXAMPLES



## BAD EXAMPLES



図 2 Caption examples generated by proposal system.

めに、画像の説明文を自動で生成する手法を提案した。従来の画像説明文生成手法が人手のラベリングを必要とするのに対して、画像・文章のみから説明文生成の学習を行える手法を設計した。具体的には、訓練サンプルの文に注目し、画像で扱われている事物とそれらの関係、さらに文法的な知識がフレーズに埋め込まれていると仮定して、そのようなフレーズを再構築することで画像説明文を生成する手法を提案した。実験では、人手による評価との相関が確かめられた各評価指標を用いて、提案手法が画像を正しく説明しうることを示した。

提案手法では、訓練サンプル文の検索精度が最終的な説明文の確からしさに直結した。今後の課題としては、この検索精度をスケラブルな範囲でより高精度にすること、文生成時に言語モデルなどの他の手掛かりを利用することが考えられる。今後の展望としては、他のモダリティを含む時系列データである動画でも同様に説明文を出力するための足掛かりになると考えられる。

## 文 献

- [1] 堀, 古井: “音声要約技術の現状とこれから”, 電子情報通信学会技術研究報告. SP, 音声, **103**, 632, pp. 21–26 (2004).
- [2] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier and D. Forsyth: “Every picture tells a story: Generating sentences from images”, ECCV, pp. 15–28 (2010).
- [3] X.-J. Wang, L. Zhang, M. Liu, Y. Li and W.-Y. Ma: “Arista - image search to annotation on billions of web photos”, CVPR, pp. 2987–2994 (2010).
- [4] K. Jung, K. I. Kim and A. K. Jain: “Text information extraction in images and video: a survey”, Pattern Recognition, **37**, 5, pp. 977–997 (2004).
- [5] B. Z. Yao, X. Yang, L. Lin, M. W. Lee and S.-C. Zhu: “I2t: Image parsing to text description”, Proc. IEEE, **98**, 8, pp. 1485–1508 (2010).
- [6] H. Nakayama, T. Harada and Y. Kuniyoshi: “Evaluation of dimensionality reduction methods for image auto-annotation”, BMVC, pp. 94.1–94.12 (2010).
- [7] Y. Ushiku, T. Harada and Y. Kuniyoshi: “Improving image similarity measures for image browsing and retrieval through latent space learning between images and long texts”, ICIP, pp. 2365–2368 (2010).
- [8] F. R. Bach and M. I. Jordan: “A probabilistic interpretation of canonical correlation analysis” (2005).
- [9] G. Salton and C. S. Yang: “On the specification of term values in automatic indexing”, Journal of documentation, **29**, 4, pp. 351–372 (1973).
- [10] F. J. Och and H. Ney: “Discriminative training and maximum entropy models for statistical machine translation”, ACL, pp. 295–302 (2002).
- [11] D. G. Lowe: “Distinctive image features from scale-invariant keypoints”, IJCV, **60**, 2, pp. 91–110 (2004).
- [12] O. Nobuyuki and T. Kurita: “A new scheme for practical flexible and intelligent vision systems”, IAPR Workshop on Computer Vision, pp. 431–435 (1988).
- [13] A. Oliva and A. Torralba: “Modeling the shape of the scene : A holistic representation of the spatial envelope”, IJCV, **42**, 3, pp. 145–175 (2001).
- [14] T. Ojala, M. Pietikäinen and D. Harwood: “A comparative study of texture measures with classification based on feature distributions”, Pattern Recognition, **29**, 1, pp. 51–59 (1996).
- [15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong: “Locality-constrained linear coding for image classification”, CVPR, pp. 3360–3367 (2010).
- [16] K. E. A. van De Sande, T. Gevers and C. G. M. Snoek: “Evaluating color descriptors for object and scene recognition.”, PAMI, **32**, 9, pp. 1582–96 (2010).
- [17] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu: “Bleu: a method for automatic evaluation of machine translation”, ACL, No. July, pp. 311–318 (2002).
- [18] G. Doddington: “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics”, HLT, pp. 138–145 (2002).