調理動画からのレシピの自動生成

牛久 敦 * 橋本 敦史 † 森 信介 ‡

概 要

画像や動画の内容を自然言語で記述する研究は大きな成果をあげ、近年注目を浴びている。有用性のあるキャプション生成という観点から、本研究では、数十分程度の調理動画を対象として、物体認識とキャプション生成を別々に行い、手順書としてレシピの生成を行う。物体認識には、Faster R-CNN を用い、その認識結果を元に、テンプレートを利用したモデルによりキャプション生成を行い、調理動画に対応するレシピとの比較を行った。生成されたレシピには、調理動画に対応するレシピと合致するようなキャプションも存在する一方、物体認識、キャプションとして出力するべきかどうかなど考慮する必要のあるキャプションも存在した。

1 序論

画像や動画といった情報と言語を結びつけることは人工知能研究における重要な問題といえる。その中でも、自動キャプション生成と呼ばれる画像や動画の内容を自然言語で自動記述するタスクは、検索への応用も踏まえ、特に重要な課題といえよう。また、検索以外にも自動キャプション生成には、内容理解補助の側面もある。例えば、動画の内容を自然言語で記述することで、実際に視聴するよりも内容理解の時間短縮ができると考えられる。しかし、既存の多くの研究は、短時間の動画の情景描写をすることに焦点を当てており、実用性が薄い。内容理解の観点からすると、動画中の単なる情景の描写ではなく、視聴者にやり方を伝えたい作業を撮影した動画から、そのやり方について言語化することは実用性が高い。

以上の点を踏まえ、我々の研究では、作業実施動画を対象とした手順書の生成を行う。作業実施動画とは、調理や機械の分解といった動作を視聴者に伝達することを意図して撮影したものであり、その内容を言語化したものを手順書という。作業実施動画から手順書を自動生成できるようになることで、視聴者の内容理解の助けになることが期待される。本研究では、KUSK Dataset [1] の調理動画を利用する。調理動画は、レシピに沿って撮影されており、この動画から元々のレシピの再現を行うことで、作業実施動画からの手順書生成の例とする。

訓練用の十分な量の動画と日本語のキャプションのデータセットは存在しないため、我々はテンプレートを用いることで、物体認識を行った結果の候補から、キャプション生成を行った。

2 関連研究

近年、Convolutional Neural Network (CNN) [2] と Long Short-Term Memmory [3] を組み合わせたモデルによる自動キャプション生成が多く研究されている [4, 5]。上記のモデルが成功を収める以前は、動画の内容として、主語、述語、目的語を確定させ、テンプレートに当てはめるキャプション生成も行われていた [6]。このようなテンプレート方式のメリットとしては、文法的に破綻したキャプションを出力しにくいことや、メディアとキャプションの組のデータを特に必要とせず、物体認識の結果をそのまま流用できる点にあった。十分な量のメディアとキャプションの組が存在する問題においては、上記の方式を利用する意味は低下したが、一方でデータセットが不十分な問題においては、画像や動画のキャプション組の新規作成は高コストである。このような場合においては、テンプレートを利用した手法も十分実用的であると考えられる。

3 提案手法

図 1 に本手法の概要を示す。まず、動画をフレームの連続として捉え CNN を利用して、各フレームごとに物体の認識を行う。次に、数フレームの物体認識の結果を用いて、コーパスから作成したテンプレートを利用することで、キャプション候補を生成する。キャプション候補にはそれぞれ、手順書としての尤もらしさを意図したスコアがつけられており、動画全体でス

*京都大学大学院 情報学研究科
†京都大学大学院 教育学研究科
‡京都大学 学術情報メディアセンター

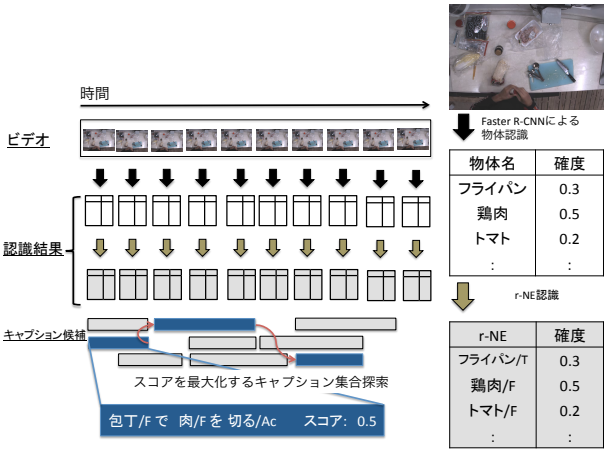

図 1: 手法概要

表 1: r-NE タグとレシピあたりの平均頻度

r-NE タグ	意味	頻度
F	食材	11.87
T	道具	3.83
D	継続時間	0.67
Q	分量	0.79
Ac	調理者の動作	13.83
Af	食材の動作	2.04
Sf	食材の状態	3.02
St	道具の状態	0.30

コアの合計値が最大になるようなキャプション列を探索し作業動画に対する手順書として出力する。

3.1 物体認識

Faster R-CNN [7] を用い、各フレームの物体認識を行う。これによって各フレーム中に存在すると推定される物体名と、物体の認識の確度を獲得する。物体認識をするタイミングは、作業内容が変更されたと考えられる調理者が物体を手に取ったり置いたりしたフレームとし、それ以外のフレームは除外した [8]。

3.1.1 レシピ固有表現

まず、我々は、調理分野特有の固有表現として レシピ固有表現 (r-NE) を定めた [9]。これは通常の固有表現と同様に単語列とタグの組で表される。特にタグについては食材や調理に関係した動作に限定している (表 1)。例えば、「チンゲン菜/F」 はチンゲン菜が食材の r-NE であることを示す。r-NE タグは固有表現認識器である PWNER [10] を利用して、認識した物体名に自動付与する。

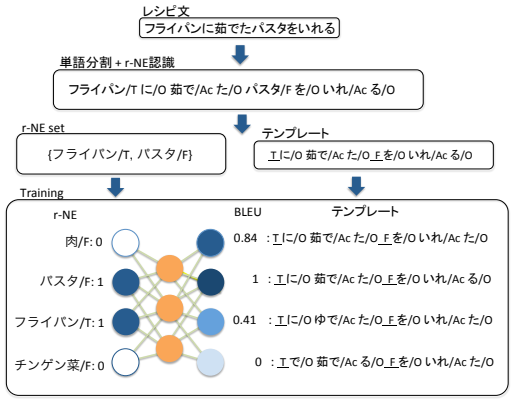

図 2: テンプレートモデルの学習. 下線が引かれている部分が、r-NE タグに置き換えられたスロットである.

3.1.2 物体認識候補

3.1 節で述べた物体認識の対象としたフレームのみを時系列順に並べ、連続する数フレームからキャプションを生成する。連続する各フレームはそれぞれ複数の r-NE 候補を持っているため、各フレームごとに 1 つずつ r-NE を選択した組み合わせを物体認識の候補として全て考える。

それぞれの組み合わせ (r-NE set) において以下の方法によりキャプション候補生成を行った。

3.2 キャプション候補生成

キャプション候補生成においては、実際のレシピ文からなるレシピコーパスを利用することで、r-NE set からキャプション候補とそのスコアの組を生成する。

3.2.1 テンプレートモデル

テンプレートモデルは、r-NE set を入力として、テンプレートごとの尤もらしさを出力する。テンプレートモデルの学習の流れを例と共に図 2 に示す。まず、前処理として、レシピコーパスの各文に対して、単語分割、r-NE 認識を行い、r-NE 部分をタグに置き換え、テンプレートを作成する。テンプレートのうち、置き換えられた r-NE 部分をスロットと呼ぶ。ただし、CNN による認識対象が物体であることを考慮し，置き換える r-NE はタグのうち F と T のみとする。また、物体認識によって情報の得ることができない r-NE がテンプレートに含まれないよう、Ac、F、T を除く r-NE タグが含まれた文は学習に使用しない。テンプレートモデルは、レシピ文の r-NE set を入力として作成さ

れたテンプレートのスコアが高くなるようにニューラルネットワークで学習を行う。

多くのテンプレートは、コーパス中に 1 度しか出現せず、1 つの r-NE set としか結びつかないため十分な学習ができないと考えられる。この問題を解決するため、BLEU [11] によって類似する文にもスコアを割り当てる。正解テンプレートをレファレンスとして全てのテンプレートの BLEU(N=4) スコアを推定する回帰問題として定式化することで、1 つの r-NE set が複数のテンプレートと結びつくことになる。

3.2.2 スコア

上記のテンプレートモデルを利用して、CNN と r-NE 認識器によって認識された r-NE set を入力としてキャプション候補とそのスコアを出力する。テンプレートのスロットに対して、r-NE を挿入することで文を生成できるが、レシピとしての尤もらしさを意図して、スコアを以下のように定める。

$$Score(S,T) = P_{tmpl}(S,T) \times \frac{1}{|S|} \sum_{s_i \in S} P_{s_i} \times P(S) \quad (1)$$

ここで、S は入力となる r-NE set、T はテンプレートを指す。$P_{tmpl(S,T)}$ は、S を入力とした時のテンプレートモデルの出力の T に対応する値を正規化したものである。これはテンプレートの尤もらしさを示す。ただし、S のタグ集合と T のスロットの対応するタグ集合が一致しない場合は、スロットに S の r-NE を挿入することができないので、値を 0 とする。s_i は、S に含まれる r-NE の 1 つであり、$|S|$ は S に含まれる r-NE の数である。$\frac{1}{|S|}\sum_{s_i \in S} P_{s_i}$ は、物体認識の結果の確度の平均を指す。これにより物体認識結果の尤もらしさの指標とする。$P(S)$ はコーパス中の S が出現する 1 文の頻度に対して $|S|$ で幾何平均をとったものである。S の要素数が多くなるほど、低頻度になることを考慮し補正を行う。これは、r-NE の組み合わせの尤もらしさを示す。

$Score(S,T)$ の値を最大化する T を探索し、そのスロットに対して、S の r-NE を挿入することでキャプション候補生成を行う。結果として、S ごとに、1 つのキャプション候補とそのスコアを得ることができる。

3.3 レシピ生成

生成したキャプション候補の集合から、合計のスコアが最大になるように、レシピに用いるキャプション集合を決定する。ただし、同一フレームの r-NE を複数回利用しない。また、同じキャプションは複数回出現しないようにする。 計算には Viterbi アルゴリズムを用い、スコアを最大化するキャプションのパスを計算することで、生成されたキャプション集合をレシピとして出力する。スコアが大きいほど、よりレシピらしいキャプション集合が出力されることが期待される。

表 2: コーパスサイズ

	文数	テンプレート数	r-NE 数
1/1	11,705	10,774	4,025
1/4	2,932	2,827	1,684
1/16	733	723	650

4 実験設定

4.1 物体認識

Faster R-CNN の学習データとして、KUSK Object Dataset [8] を用いた。これは、レシピに対応する調理作業の動画を 3 つのカメラで撮影したものに対して、食材や調理器具についてアノテーションを行ったものである。また、1 つのレシピデータに対して、別の人間が行った調理動画が存在する。

ある動画に対してレシピを生成する際には，その動画から得られたデータは取り除いて学習したモデルを用いて物体認識を行った。認識結果候補の r-NE set 作成時のフレーム幅は 3 とした。

4.2 テンプレートモデル

表 2 に、テンプレートモデルの学習データの設定を示す。レシピコーパスは以下の条件を満たすレシピ文の集合である。

- F か T の r-NE が 1 個以上 3 個以下含まれる
- Ac、F、T 以外の r-NE タグが含まれていない

r-NE set を入力として、$P_{tmpl}(S,T)$ を計算するテンプレートモデルとして 3 層ニューラルネットワークを構成する。epoch 数は 200 で、バッチサイズは 50、中間層のユニット数は 1,000 に設定した。最適化関数としては Adam [12] を利用した。また、誤差関数を 2 乗誤差に設定した。

4.3 評価

コーパスをそのまま用いたものに加え、1/4, 1/16 のサイズのコーパスによってテンプレートモデルを作

表 3: 実験結果

	BLEU			
サイズ	N=1	N=2	N=3	N=4
1/1	27.32	14.98	7.97	3.55
1/4	30.92	**17.18**	**9.13**	**5.08**
1/16	**31.19**	16.02	7.96	2.99

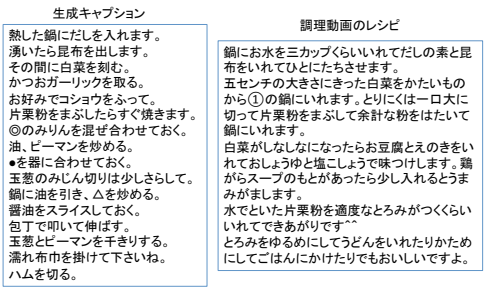

図 3: 生成されたレシピと調理動画のレシピの例

成した。評価方法として、調理動画のレシピと前章の手法により生成されたレシピを BLEU (N=1-4) を用いて評価した。今回は KUSK Object Dataset のうち 7 個のレシピに対応する 16 個の動画に対して実験を行った。

5 実験結果

実験結果を表 3 に示す。学習のサイズに関しては 1/4 の時が最も良い結果になった。

5.1 考察

図 3 の実際に生成されたレシピと、調理動画のレシピの例を示す。内容として違和感のないキャプションも複数生成されているのが分かる。

一方で、生成されたレシピには、明確に生成するべきでないキャプションも存在した。このようなキャプションが生成される理由としては、物体認識誤りや、キャプションとするべきでないものについて言及してしまっていることなどが挙げられる。例えば、「ハムを切る」は、動画中に存在していない「ハム」を誤認識してしまっている。また、「濡れ布巾を掛けて下さいね」は、本来キャプションとすべきでない「布巾」について言及している例である。これらの問題を解決することで、より適切なレシピ生成ができると考えられる。

6 結論

作業実施動画からの手順書の生成の例として、調理動画に対して、CNN とテンプレートを利用することで、一定の精度でレシピの生成を可能にした。

一方で、生成されたレシピは、調理動画の内容を正しく記述できていない部分も存在する。物体認識の精度向上や認識された r-NE set がキャプションに用いるべきかどうかについて考慮することでより適切なレシピ生成が行えると期待される。

7 謝辞

本研究は JSPS 科研費 26280084 の助成を受けたものである。ここに謝意を表する。

参考文献

[1] Hashimoto, A., Tetsuro, S., Yamakata, Y., Mori, S. and Minoh, M.: KUSK Dataset: Toward a Direct Understanding of Recipe Text and Human Cooking Activity, *Proc. of CEA* (2014).

[2] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proc. of IEEE*, Vol. 86, No. 11 (1998).

[3] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8 (1997).

[4] Guo, Z., Gao, L., Song, J., Xu, X., Shao, J. and Shen, H. T.: Attention-based LSTM with Semantic Consistency for Videos Captioning, *Proc. of ACMMM*, ACM (2016).

[5] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K.: Sequence to sequence-video to text, *Proc. of ICCV* (2015).

[6] Krishnamoorthy, N., Malkarnenkar, G., Mooney, R. J., Saenko, K. and Guadarrama, S.: Generating Natural-Language Video Descriptions Using Text-Mined Knowledge., *AAAI*, Vol. 1 (2013).

[7] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems* (2015).

[8] Hashimoto, A., Mori, S., Iiyama, M. and Minoh, M.: KUSK Object Dataset: Recording Access to Objects in Food Preparation, *Proc. of ICME*, IEEE (2016).

[9] 笹田鉄郎, 森信介, 山肩洋子, 前田浩邦, 河原達也: レシピ用語の定義とその自動認識のためのタグ付与コーパスの構築, 自然言語処理 (2015).

[10] Sasada, T., Mori, S., Kawahara, T. and Yamakata, Y.: Named Entity Recognizer Trainable from Partially Annotated Data, *Proc. of PACLING*, ACM (2015).

[11] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proc. of ACL* (2002).

[12] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).