リカレント連続翻訳モデル

Nal KalchbrennerPhil Blunsom
Department of Computer Science
University of Oxford
{nal.kalchbrenner,phil.blunsom}@cs.ox.ac.uk

アブストラクト

我々は、Recur- rent Continuous Translation Modelsと呼ばれる確率的連続翻訳モデルのクラスを紹介する。このモデルは、純粋に単語、フレーズ、文の連続的な表現に基づいており、アライメントやフレーズ翻訳ユニットには依存しない。このモデルには、生成の側面と処理の側面があります。翻訳の生成はターゲットリカレントランゲージモデルでモデル化され、原文の条件付けはConvolu-tional Sentence Modelでモデル化される。様々な実験を通して、我々のモデルは、金の翻訳に対するパープレキシティが、最新のアライメントベースの翻訳モデルに比べて43%*以上低いこと*を示す。次に、アラインメントがないにもかかわらず、原文の語順、構文、意味に非常に敏感であることを示す。最後に、*n*-bestの翻訳リストを再スコアリングする際に、最先端のシステムと一致することを示す。

1 はじめに

機械翻訳の統計的アプローチでは、翻訳の基本単位は1つまたは複数の単語からなるフレーズです。翻訳システムの重要な構成要素は、ソース言語とターゲット言語のフレーズのペアの翻訳確率を推定するモデルです。このようなモデルでは、フレーズの表面上の形が異なる場合、フレーズのペアとその出現頻度を区別して数えます。別々のフレーズペアは、しばしば重要な類似性を持っていますが

言語的なものであれ、その他のものであれ、それらはモデルの変換確率の推定において統計的な重みを共有していません。フレーズペアの類似性を無視することに加えて、これは一般的なスパース性の問題につながる。フレーズの長さに比例して指数関数的に増加する多数の稀なフレーズペアや見たことのないフレーズペアでは、推定値が疎かになったり歪んだりし、他のドメインへの一般化には限界があることが多い。

このような問題に対処するために、連続的な表現が期待されています。単語の連続表現は、形態的、構文的、意味的な類似性を捉えることができます(Collobert and We-ston, 2008)。連続的な言語モデルに適用され、スパース性の問題を克服し、最先端の性能を達成することができました(Bengio et al., 2003; Mikolov et al., 2010)。また，単語表現は，条件付け情報に対して顕著な感度を示している(Mikolov and Zweig, 2012)．文字の連続表現は，文字レベルの言語モデルに導入され，表を作らない言語生成能力を示した(Sutskever et al., 2011)．また，フレーズやセンテンスについても，連続的な表現が構築されている．これらの表現は、類似性やタスクに依存する情報（例えば、感情、言い換え、対話のラベルなど）を、単語レベルを大幅に超えて伝達することができ、非常に多様な未見のフレーズやセンテンスのラベルを正確に予測することができる(Grefenstette et al., 2011; Socher et al., 2011; Socher et al., 2012; Hermann and Blunsom, 2013; Kalchbrenner and Blunsom, 2013)。

フレーズベースの連続翻訳モデルは、(Schwenk et al, 2006)で最初に提案され、再

自然言語処理における経験的手法に関する2013年会議の議事録、ページ1700～1709、米国ワシントン州シアトル、2013年10月18～21日。Qc 2013 計算言語学協会

最近では(Schwenk, 2012; Le et al., 2012)でさらに発展させました。このモデルには、翻訳確率を推定する原理的な方法が組み込まれており、希少なフレーズや見たことのないフレーズにもロバストに対応しています。これらのモデルは、Bleuスコアを大幅に向上させ、より示唆に富む翻訳を実現しています。しかし、これらのモデルは、固定サイズのソースおよびターゲットフレーズに限定されており、制限されたターゲット言語モデリング情報を考慮して、ターゲットワード間の依存関係をシミュレートしています。

本稿では、リカレント連続翻訳モデル（Recurrent Continuous Translation Models: RCTM）と呼ばれる連続翻訳モデルについて述べる。我々は2つのRCTMアーキテクチャを定義した。いずれのモデルも、目標とする翻訳の生成にリカレント言語モデルを採用しています(Mikolov et al., 2010)。他のn-gramアプローチとは対照的に、リカレント言語モデルは、ターゲット文の単語の依存関係についてマルコフ仮定をしません。

2つのRCTMは、ソース文に対するターゲット言語モデルのコンディジョン方法が異なります。1つ目のRCTMは、畳み込みセンテンスモデル(Kalchbrenner and Blunsom, 2013)を用いて、ソースの単語表現をソース文の表現に変換します。この原文の表現が、各ターゲット単語の生成を制約します。第2のRCTMでは、中間的な表現を導入しています。これは、畳み込み文モデルのトランケートされた変形を使用して、まず原語の表現を目標語の表現に変換し、目標語の表現が目標文の生成を制約します。いずれの場合も、畳み込み層は、文の中の単語の表現から、文の中のフレーズの組み合わせ表現を生成するために使用されます。

RCTMの利点は、潜在的なアライメント・セグメンテーションがないことと、それに関連するスパース性です。ソースとターゲットの単語、フレーズ、センテンスの間のつながりは、それらの連続的な表現の間のマッピングとして暗黙のうちにのみ学習されます。セクション5で見たように、これらのマッピングは、次のようになります。

これらのモデルの下での翻訳の確率は、原文と訳文の長さに線形な少数の行列-ベクトル積を必要とする効率的な計算が可能であること。さらに、RCTMの確率分布から直接翻訳を生成することができ、外部リソースを必要としません。

4つの実験でモデルの性能を評価した。RCTMの翻訳確率は扱いやすいので、参照翻訳に対するモデルのパープレキシティを測定することができる。モデルの複雑さは、IBMモデル1よりも有意に低く、IBMモデル2の最先端のバリエーションの複雑さよりも43 % *以上*低い（Brown et al.1993; Dyer et al.2013）。2番目と3番目の実験は、RCTM IIの出力が原文の言語情報に対してどのような感度を持つかを示すことを目的としています。2つ目の実験では、原文の単語をランダムに並べ替えると、参照元の翻訳に対するモデルの困惑度が大幅に悪化することを示し、このモデルが単語の位置と順序に非常に敏感であることを示唆している。第3の実験では、RCTM IIが生成した翻訳を検証した。生成された翻訳は、原文と形態素、構文、意味の面で顕著な一致を示した。 最後に、*n*-bestリストの翻訳を再スコアリングするタスクでRCTMをテストする。RCTMの確率に1つの単語ペナルティ機能を加えたものは、5つのアライメントベースの翻訳モデルを含む12の機能を利用した最先端の翻訳システムcdecの性能に匹敵する(Dyer et al., 2010)。

以下のように進めます。まず第2章では、RCTMを支える一般的なモデリングの枠組みについて説明します。セクション3では、RCTM Iを説明します。3節ではRCTM I、4節ではRCTM IIについて説明します。4ではRCTM IIを説明します。セクション5では、4つの実験について説明し、セクション6で結論を述べます。[1]

2 フレームワーク

まず、RCTMのモデリングフレームワークについて説明します。RCTMは、原文$e = e_1, ..., e_k$の翻訳である標的文$f = f_1, ..., f_m$の確率$P(f|e)$を推定する。ここで

10は、非常に正確な形態学的、シンタックス学的な知識 を持っています。 ティックな情報とセマンティックな情報を得ることができます。もう一つの利点は

[1]コードとモデルはnal.coで入手可能

ここでは、単語$f_i, ..., f_j$の部分文字列を$\mathbf{f}_{i:j}$で表します。次のような恒等式があります。

$$P(\mathbf{f}|\mathbf{e}) = \prod_{i=1}^{m} P(\mathbf{f}_i | \mathbf{f}_{1:i-1}, \mathbf{e}) \quad (1)$$

RCTMは、$P(\mathbf{f}|\mathbf{e})$を直接計算して推定します。各ターゲットの位置iに対して、条件付確率の中でターゲットワード$\mathbf{f}_i$が出現する*確率*$P(\mathbf{f}_i | \mathbf{f}_{1:i-1}, \mathbf{e})$。

RCTMは、原文$\mathbf{e}$だけでなく、目標文$\mathbf{f}_{1:i-1}$の先行する単語$\mathbf{f}$にも敏感に反応することで、目標言語そのもののモデルを組み込んでいることがわかります。

条件付き確率$P(\mathbf{f}|\mathbf{e})$をモデル化するために、RCTMは、ターゲット文の生成アーキテクチャと、ソース文にターゲット文を条件付けするアーキテクチャの両方から構成されます。式(1)を完全に表現するために、生成アーキテクチャを、再帰ニューラルネットワークに基づく再帰言語モデル(RLM)でモデル化します(Mikolov et al., 2010)。RLMにおけるi番目の単語$\mathbf{f}$の予測$_i$は、対象文の先行するすべての単語$\mathbf{f}$に依存する$_{1:i-1}$ため、式1に条件付き独立性の仮定が導入されない。予測は、$\mathbf{f}$の$_i$直前の単語に最も強く影響されますが、文全体からの長距離依存性を示すこともあります。予測のアーキテクチャはモデルに依存しており、セクション3-4で扱う。3-4.モデルの生成面と構成面の両方とも、構成要素のための連続的な表現を展開し、単一の共同アーキテクチャとして学習されます。RCTMの基礎となるモデリング・フレームワークを踏まえて、生成的な側面に基づくリカレント言語モデルの詳細を説明します。

2.1 リカレント言語モデル

RLMは、与えられた言語の中で、ある単語$\mathbf{f}$が出現する確率$P(\mathbf{f})$をモデル化します。ここで$\mathbf{f} = \mathbf{f}_1, ..., \mathbf{f}_m$は、$m$*個の*単語のシーケンスで、ターゲット言語のセンテンスなどである。式(1)と同様にアイデンティティを使って

$$P(\mathbf{f}) = \prod_{i=1}^{m} P(\mathbf{f}_i | \mathbf{f}_{1:i-1}) \quad (2)$$

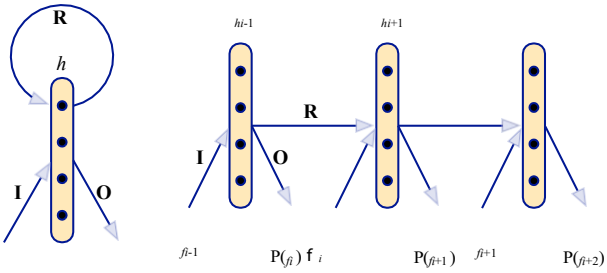

図1：RLM（左）とその深さ3までの解像（右）。リカレント変換は隠蔽層$\boldsymbol{h}$に適用され$_{i-1}$、その結果は現在の単語$\mathbf{f}$の$_i$表現に合計されます。非線形変換の後、次の単語$\mathbf{f}$に対する$_{i+1}$確率分布が予測される。

$P(\mathbf{f}_i | \mathbf{f}_{1:i-1})$である。RLMのアーキテクチャは、言語の単語$\mathbf{f}$を含む$_i$語彙$V$と、3つの変換から構成されている。電流変換$\mathbf{R} \in \mathbb{R}^{q \times q}$と出力語彙変換$\mathbf{O} \in \mathbb{R}^{|V| \times q}$。各単語$f_k \in V$に対して、$V$におけるインデックスを$\mathbf{i}(f_k)$で示し、$\mathbf{v}(\mathbf{f}_k) \in \mathbb{R}^{|V| \times 1}$で、$\mathbf{v}(\mathbf{f}_k)_{\mathbf{i}(f_k)}=1$のみのオールゼロベクトルを示します。

ある単語$\mathbf{f}$に対して$_i$、$\mathbf{I}\,\mathbf{v}(\mathbf{f}_i) \in \mathbb{R}^{q \times 1}$の結果は$\boldsymbol{h}_i$の入力連続表現である$_i$. パラメーター$q$は、単語表現のサイズを決定します。予測は、リカレント変換$\mathbf{R}$を単語表現に連続的に適用し、各ステップで次の単語を予測することで進みます。詳細には、各$P(\mathbf{f}_i | \mathbf{f}_{1:i-1})$の計算は再帰的に進みます。$1 < i < m$の場合。

$$\boldsymbol{h}_1 = \sigma(\mathbf{I} \cdot \mathbf{v}(\mathbf{f}_1)) \quad (3a)$$

$$\boldsymbol{h}_{i+1} = \sigma(\mathbf{R} \cdot \boldsymbol{h}_i + \mathbf{I} \cdot \mathbf{v}(\mathbf{f}_{i+1})) \quad (3b)$$

$$\boldsymbol{o}_{i+1} = \mathbf{O} \cdot \boldsymbol{h}_i \quad (3c)$$

であり、条件付き分布は次のように与えられる。

$$P(\mathbf{F}_i = V | \mathbf{F}_{1:i-1}) = \frac{\exp(o_{i,v})}{\sum_{v=1}^{V} \exp(o_{i,v})} \quad (4)$$

式（3）において、σはtanhなどの非線形関数です。バイアス

の値b_hとb_oが計算に含まれます。となります。

このモデルは、条件付き分布を単純な仮定なしに明示的に計算します。

図1にRLMの説明図を示します。

RLMの学習は、バックプロパゲーションにより

時間である(Mikolov et al.,
2010)。出力層で計算された事前予測された分布の誤差は

を再帰層でバックプロパゲーションし、与えられたステップ数dの前回の予測の誤差に加算します。この手順は、図1のように深さdまで解かれたRLMに対する標準的なバックプロパゲーションと同等である。

RCTMは、各単語fの予測分布$_i$が原文eに制約されているRLMと考えることができます。

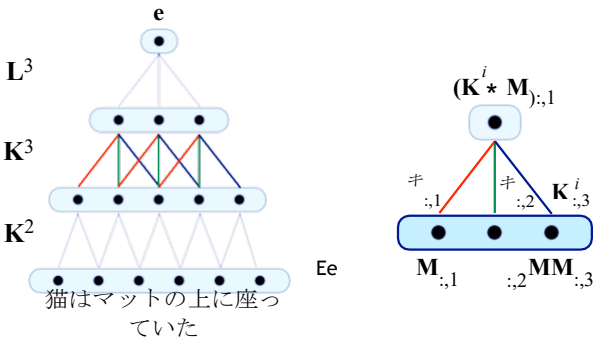

図2：6語の原文eと計算された文表現**e**のCSM。**K**^{2}, **K**^{3}は重み行列、**L**^{3}はトップの重み行列。右側は、ある重み行列**K**^{i}と、例えば**E**に対応するe可能性のある一般的な行列**M**との間の一次元畳み込みの例である。重みの色分けは、重みの共有を示しています。

3　リカレント連続翻訳モデルI

RCTM Iでは、条件付けに畳み込み文モデル（Convolutional Sentence Model: CSM）を採用している。CSMは、文の中の*n*-gramの表現から段階的に構築される文の表現を作成する。CSMは階層的な構造を持つ。表現を生成する演算は、明示的な解析木を使用していませんが、モデルの下層では小さな*n-gram*に対して局所的に作用し、モデルの上層では文全体に対してよりグローバルに作用するようになっています。構文木を必要としないことは、構文木を必要とする文モデルと比較して、2つの主要な利点をもたらす(Grefenstette et al., 2011; Socher et al., 2012)。1つ目は、正確なパーサーが利用できない多くの言語にモデルをロバストに適用できることです。第二に、ターゲットセンテンス上の翻訳確率分布は、選択された構文解析ツリーに依存しません。

RCTM Iは、CSMによって生成された原文eの連続的な表現に基づいて、各$_i$ターゲット単語fの確率を条件付けます。これは、ターゲット再帰言語モデルの各隠れ層hに$_i$文の表現を追加することで達成される。次に、CSM自体から始めて、手順をより詳細に説明します。

3.1　畳み込み文モデル

CSMは、文の連続表現を、連続表現に基づいてモデル化します。

は、文中の単語のe = e$_1$ ... e$_k$をある言語の文とし、v(e $_i$) $\in^{Rq\times 1}$を単語 e $_i$の連続表現とする。

また、**E**$^e\in^{Rq\times k}$を次のように定義されるeの*文行列*とする。

CSMのアーキテクチャの主な構成要素は、一連の*重み*行列（**K**^{i}）であり$_{2\le i\le r}$、これは畳み込みのカーネルまたはフィルタに対応しており、学習された特徴検出器と考えることができる。CSMは、文行列**E**から、重みe行列によって与えられる重みを持つ一連の畳み込みを**E**にe適用することで、文eの連続ベクトル表現**e** $^{Rq\times 1}$を計算する。重み行列と畳み込みのシーケンスを次に定義する。

ここで$_{2\le i\le r}$, **K** $^{iRq\times i}$ は i 列の行列であり，r =∈ $2N$ とすると，N は学習セットの中で最も長い原文の長さである．**K**の^{i}各行はi*個の*重みのベクトルで，*1次元*の畳み込みのカーネルまたはフィルタとして扱われます。例えば，列数が j i である行列 **M** $^{Rq\times j}$ が与えられた場合，**K** ^{i}の各行は **M** の対応する行と畳み込むことができ，結果として

ここで，∗は畳み込み操作を示し，（**K**^{i} ∗ **M**）$\in^{Rq\times(j-i+1)}$とします^i。i = 3の場合、値（**K**^{i} ∗ **M**）は$_{:,a}$次のように計算されます。

$$\mathbf{K}^i_{:,1}\ \mathbf{M}_{:,a} + \mathbf{K}_{:,2}\ \mathbf{M}_{:,a+1} + {}^{Ki}_{:,3}\ \mathbf{M}_{:,a+2} \quad (6)$$

*列*とする。

ここでは、成分単位のベクトル積です。 Applying the convolution kernel $\mathbf{K}^i$ yields a matrix $(\mathbf{K}^i \mathbf{M})$ that has i 1 columns less than the original matrix $\mathbf{M}$.

$$\mathbf{E}^{e}_{:,i} = \mathsf{V}(\mathsf{E}_i) \quad (5)$$

長さkの原文が与えられると、CSMは

*–

は、文の行列$\mathbf{E}$と連続して畳み込みます。 e

は、次のように$\mathbf{K}$から^{2}始まる一連の重み行列（$\mathbf{K}^i$）を$_{2\le i\le r}$、1つは他のものと重ね合わせます。

$$\mathbf{E}^{e}_{1} = \mathbf{E}^{e} \quad (7a)$$

$$\mathbf{E}^{e}_{i+1} = \sigma(\mathbf{K}^{i+1} * \mathbf{E}^{e}_{i}) \quad (7b)$$

数回の畳み込み操作の後，$\mathbf{E}^{e}_{i}$は$^{Rq\times 1}$のベクトルとなり，この場合は目的の表現が得られたことになるが，$\mathbf{E}^{e}$の$_{i}$列数が次の重み行列$\mathbf{K}^{i+1}$の列数$i+1$よりも小さい場合もある．後者の場合は，$\mathbf{E}^{e}$と同じ列数を持つ上位の重み行列$\mathbf{L}^{j}$を適用するだけで，同様に$^{Rq\times 1}$のベクトルが得られる．このようにして，原文eの文表現$\mathbf{e}_{i}$
$^{Rq\times 1}$が得られる．式7bの畳み込み演算には，非線形関数σが挿入されていることに注意してほしい．また，重み行列$\mathbf{K}^i$と$\mathbf{L}^i$が適用されるレベルが異なる場合，トップの重み行列$\mathbf{L}^j$は，（$\mathbf{K}^i$）$_{2\le i\le r}$$^{2\le i\le r}$とは異なる重み行列（$\mathbf{L}^i$）の追加シーケンスから得られることにも注意してください．図2は，CSMと1次元の畳み込みの例を示したものである．^{2}

3.2 RCTM I

RCTM Iは、第2節で定義したように、ターゲット言語Fの文$\mathsf{f} = \mathsf{f}_1, ..., \mathsf{f}_m$がソース言語Eの文$\mathsf{e} = \mathsf{e}_1, ..., \mathsf{e}$の$_k$翻訳であるという条件付き確率$P(\mathsf{f}|\mathsf{e})$をモデル化し、式1に従って条件付き分布$P(\mathsf{f}_i|\mathsf{f}_{1:i-1}, \mathsf{e})$を明示的に計算する。RCTM Iの構造は、ソース言語V^{E}とターゲット言語V^{F}、構成要素であるCSMの2つの重み行列（$\mathbf{K}^i$）$_{2\le i\le r}$と（$\mathbf{L}^i$）$_{2\le i\le r}$、構成要素であるRLMの変換$\mathbf{I}$ $^{Rq\times|V^{\mathsf{F}}|}$、$\mathbf{R}$ $^{Rq\times q}$、$\mathbf{O}$ $^{|V^{\mathsf{F}}|\times q}$、文変換$\mathbf{S}$ $^{Rq\times q}$から構成される。eを入力文とするCSMの出力を$\mathbf{e} = \mathsf{csm}(\mathsf{e})$とする。

RCTM Iの計算は、以下で説明したRLMの計算を簡単に変更したものです。

Eq.3.次のように再帰的に進みます。

$$\mathbf{s} = \mathbf{S} - \mathsf{csm}(\mathsf{e}) \quad (8a)$$

$$h_1 = \sigma(\mathbf{I} - \mathsf{v}(\mathsf{f}_1) + \mathbf{s}) \quad (8b)$$

$$h_{i+1} = \sigma(\mathbf{R} \cdot h_i + \mathbf{I} \cdot \mathsf{v}(\mathsf{f}_{i+1}) + \mathbf{s}) \quad (8c)$$

$$o_{i+1} = \mathbf{O} \cdot h_i \quad (8d)$$

σは非線形関数であり，計算にはバイアス値が含まれています。図3はRCTM I.の例です。

RCTM Iの2つの側面に注目してください。まず、ターゲット文の長さは、ターゲットRLM自身によって予測されますが、RLMはそのアーキテクチャ上、短い文に偏っています。第二に、原文の表現は、すべての目標語を一様に制約する。これは、目標語が原文の特定の部分に強く依存し、他の部分にはあまり依存しないという事実に反している。次のモデルでは、これらの側面を別の形で表現することを提案する。

4 リカレント連続翻訳モデルII

RCTM IIの中心的なアイデアは、まず、主要なアーキテクチャとは無関係に、ターゲットセンテンスの長さmを推定することです。mと原文eが与えられると、モデルはeのn-gramの表現を構築します（nは4に設定）。したがって、eの4-gram表現は、$n = 4$に対応するレベルでCSMをトランケートすることで構築されます。次に、この手順を逆にします。ソース文eの4-gram表現から、モデルはターゲットの予測される長さmを持つ文の表現を構築する。これも同様に、長さmの文に対して*反転した*CSMを切り捨てることで達成される。

次に、CGM（Convolutional n-gram Model）の詳細を説明します。その後、RCTM IIの説明に戻ります。

4.1 畳み込み式N-gramモデル

CGMは、選択されたnの値に対してn-gramが表現されているレベルでCSMを切り捨てて得られます。行列$\mathbf{E}$の^{e}列$\mathbf{g}_i$は式（7）で表される$_{i}$からのn-gramです。

^{2}この構造の正式な取り扱いについては、（Kalchbrenner and Blunsom, 2013）を参照してください。

nの値は、n-gram表現$\mathbf{g}$ が構築される単語ベクトルの数に対応し、等価的に、nは$\mathbf{g}$ の下にあるCSMの重みのスパンである（図2-3参照）。なお、行列$\mathbf{E}$のどの列も、同じスパン値のn-gram^{e}を表しています n.n-gramのサイズを$\mathsf{gram}(\mathbf{E}^{e})$とします。

i

i

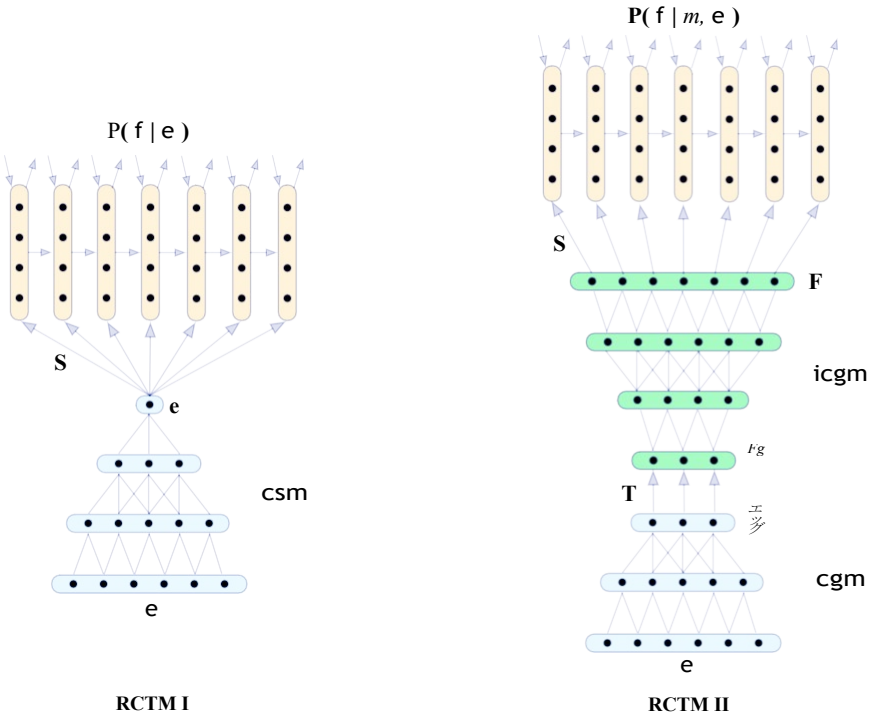

図3：2つのRCTMを図式化したもの。矢印は完全な行列変換を表し，線は重み行列の列に対応するベクトル変換を表す.

E
^{e}で表される。例えば、$_{2}$十分に長い文eの場合、gram(**E** ^{e}) = 2, gram(**E** ^{e}) = 4, gram(**E** ^{e}) = 7となる。ここで、原文eのn-gramを表すCSM^{e}の行列**E**をcgm(e, n)とする。

CGMを反転させることで、文のn-gramの表現から文の表現を得ることもできる。これは、n-gram表現cgm(e, n)のサイズと対象文の長さに依存する。

m. 変換icgmはn-gramを展開します。

表現を、ターゲットとなる感覚の表現に置き換えています。

はm 個 の単語で構成されています。このアーキテクチャは、逆CGM、あるいは同等に、逆切断CSMに対応しています（図3）。cgmとicgm の 変 換を受けて、RCTM
IIの計算を詳しく説明します。

4.2 RCTM II

RCTM
IIは、条件付き確率をモデル化しています。

を計算し、分布P (f $_{i+1}$f$_{1:i}$, m, e)とP (m e)を算出します。RCTM |
IIのアーキテクチャは、RCTM
Iのすべての要素に加えて、以下の追加要素から構成されています。
並進変換$\mathbf{T}^{q\times q}$と、icgmの一部である^{3}2つの重み行列 （$\mathbf{J}^{i}$） $_{2\le i\le s}$および （$\mathbf{H}^{i}$） $_{2\le i\le s}$のシーケンスです。

RCTM
IIの計算は、以下のように繰り返し行われます。

$$\mathbf{E}^{g} = \mathrm{cgm}(e, 4^{\vdots,}_{j}) \quad (10a)$$

$$\mathbf{F}^{q}_{\vdots,\,g} = \sigma(\mathbf{T} - \mathbf{E}) \quad (10b)$$

P (f|e) を次のようにファクタリングします。

$$P(f|e) = P(f|m, e) - P(m|e) \quad (9a)$$

$$= \sum_{i=1}^{m} P(f_{i+1}|f_{1:i}, m, e) - P(m|e) \quad (9b)$$

$$\mathbf{F} = \text{icgm}(\mathbf{F}^{g}, m) \quad (10c)$$

$$h_1 = \sigma(\mathbf{I} - \text{v}(\text{f}_1) + \mathbf{S} - \mathbf{F}_{:,1}) \quad (10d)$$

$$h_{i+1} = \sigma(\mathbf{R} \cdot h_i + \mathbf{I} \cdot \text{v}(\text{f}_{i+1}) + \mathbf{S} \cdot \mathbf{F}_{:,i+1}) \quad (10e)$$

$$o_{i+1} = \mathbf{O} \cdot h_i \quad (10f)$$

とし、条件付き分布$P(\text{f}_{i+1} \text{f}_{1:i}, \text{e})$を式4のように$o_i$から求める。再構成された各ベクトル$\mathbf{F}_{:,i}$は、対象となる単語$\text{f}_i$を予測する対応する層$h$に順次追加されていくことに注意してください。RCTM IIの構成を図3に示します。

[3] rと同様に、値sは小さく、トレーニングセットのソース文とターゲット文の長さに依存します。5.1.2節参照。

翻訳の長さを個別に推定するために、条件付き確率$P(m|\mathbf{e})$を次のようにして推定します。

$$P(m|\mathbf{e}) = P(m|k) = \text{Poisson}(\lambda_k) \quad (11)$$

ここで、kは原文$\mathbf{e}$の長さであり$\text{Poisson}(\lambda)$は平均λのポアソン分布です。()
以上でRCTM
IIの説明を終わります。次に、実験について説明します。

5 実験の様子

4つの実験について報告します。最初の実験では、参照翻訳に対するモデルのパープレクシティを検討した。2回目と3回目の実験では、原文の言語的側面に対するRCTM
IIの感度をテストした。最後の実験では、2つのモデルの再スコアリングの性能を検証する。

5.1 トレーニング

実験に入る前に、RCTMの学習に使用したデータセット、ハイパーパラメータ、最適化アルゴリズムについて説明します。

5.1.1 データセット

実験に使用したトレーニングセットは，Eighth Workshop on Ma-chine Translation (WMT) 2013
のトレーニングデータのうち，ニュースの解説部分に含まれる，長さが80語以下のセンテンスのペア144953対からなるバイリンガルコーパスである。ソース言語は英語で、ターゲット言語はフランス語である。英語の文には約410万語、フランス語の文には約450万語の単語が含まれています。英語文とフランス語文の両方で、2回以下しか出現しない単語は、*未知の*トークンで置き換えられます。結果として得られた語彙Vと$^{E}V^{F}$には、それぞれ25403個の英語の単語と34831個のフランス語の単語が含まれています。

実験では，2009年，2010年，2011年，2012年の Workshop on Machine Transla-tion News Test (WMT-NT)
セットからなる4種類のテストセットを使用した。それぞれのテストセットの内容は以下の通りです。
2525, 2489, 3003, 3003組の英仏文を用意しました。これらのデータセットに含まれる未知の単語は、*未知の*トークンに置き換えて実験を行いました。また、2008年のWMT-
NTセットでは、2051組の英仏文が収録されており、これを検証セットとして使用した。

5.1.2 モデルのハイパーパラメータ

e $_i V$ に対する En-glish vector v(e$_i$) のサイズE, hid-den 層 h_i のサイズ，および v(f$_i$) V に対する French vector v(f$_i$) のサイズを定義するパラメータ q は，q = 256 にF設定されています．これにより，比較的小さなリカレント行列とそれに対応するモデルが得られる．学習を高速化するために，(Mikolov et al., 2011)の手順に従って，ターゲット語彙 V をF256 のクラスに因数分解する．

RCTM IIでは、nを4に設定した畳み込みn-gramモデルCGMを使用しています。RCTM Iでは、CSMの重み行列の数r は15であるのに対し、RCTM IIでは、CGMの重み行列の数r は7、逆CGMの重み行列の数sは9である。テスト文がすべての訓練文よりも長く、モデルに大きな重み行列が必要な場合、大きな重み行列は、重みが訓練された2つの小さな重み行列に容易に因数分解される。例えば、10個の重みの行列が必要だが、重み行列が重み9までしか学習されていない場合、10個の重みの行列を、9と2の1つずつで因数分解することができる。

5.1.3 目的と最適化

目的関数は、フランス語の文の中の予測された単語と真の単語のクロスエントロピーエラーの合計の平均です。英語の文は、フランス語の文を予測する際の入力として用いられますが、それ自体は予測されません。目的語には正則化項 (l) $_2$が追加されます。モデルの学習は、時間を通じたバックプロパゲーションによって行われます。各ステップの出力層で計算されたクロス・エントロピー・エラーは、リカレント構造を介して、ステップ数dだけ逆伝播されますが、すべてのモデルでd = 6としました。隠れ層に蓄積された誤差は、変換**S**とCSM/CGMを介して、英語の入力文e の入力ベクトルv(e $_i$)にさらに逆伝播される。英語のベクトルを含むすべての重みは、ランダムに初期化され、学習中に推測される。

目的語はミニバッチ適応勾配降下法（Adagrad）を用いて最小化されます(Duchi et al., 2011)。RCTMの学習には、3台のマルチコアCPUで約15時間かかります。我々の実験では

∈

∈

WMT-NT	*2009*	*2010*	*2011*	*2012*
KN-5	218	213	222	225
RLM	178	169	178	181
IBM 1	207	200	188	197
FA-IBM 2	153	146	135	144
RCTM I	143	134	140	142
RCTM II	**86**	**77**	**76**	**77**

表1：WMT-NTセットでのPerplexityの結果。

は比較的小さいですが、原理的には、我々のモデルは何億もの単語に適用されたRLMと同様にスケールすることができると考えています。

5.2 金の翻訳の戸惑い

あるRCTMの下での翻訳の確率の計算は効率的なので、テストセットの参照翻訳に関するRCTMの錯綜度を計算することができます。perplexityは、モデルが翻訳に割り当てる品質を示す指標である。RCTMのパープレクシティを、IBM Model 1 (Brown et al., 1993)およびIBM Model 2の最新版であるFast-Aligner (FA-IBM 2)モデルのパープレクシティと比較した(Dyer et al., 2013)。ベースラインとして，無条件のターゲットRLMと，修正Kneser-Nayスムージングを用いた5グラムのターゲット言語モデル（KN-5）を追加した。結果はTab.1で報告されています。1.RCTM IIは，以下のようなパープレキシティを得ることができました.
>
RCTMのパープレキシティが低いのは、連続的な表現とそれらの間の変換が、明示的なアラインメントの欠如をうまく補っていることを示唆している。さらに、RCTM自体のperplexityの違いは、条件付けアーキテクチャの重要性を示しており、RCTM IIの局所的な4-gram条件付けは、RCTM Iの原文全体を使った条件付けよりも優れていることを示唆している。

5.3 原文の文構造に対する感度

2つ目の実験は，英語原文中の単語の順序と位置に対するRCTM IIの感度を示すことを目的としている。この目的のために，トレーニングセットとテストセットをランダムに入れ替えた。

wmt-nt perm	*2009201020112012*
RCTM II	174168175178

表2：英語原文の単語をランダムに並べ替えたWMT-NTセットに対するRCTM IIのPerplexityの結果。

は，英語の原文に含まれている単語のリストです。並べ替えを行ったデータの結果を表2に示します。2.もし、RCTM IIがBag-of-Wordsアプローチとほぼ同等であれば、単語の並べ替えによる違いはないと考えられます。一方、Tab.2で報告された結果と、Tab.2で報告された結果の違いは、以下のとおりです。2とTab.1で報告された結果の差は非常に大きく、明らかにこれは、翻訳モデルが語順と位置に敏感であることを明確に示しています。

5.3.1 RCTM IIからの生成

RCTM IIが語順だけでなく、文の他の統語的・意味的特徴にも敏感であることを示すために、様々な英語の原文に対して訳語を生成し、検査を行った。この生成は、RCTM II自身の確率分布からのサンプリングによって行われ、他の外部リソースには依存しない。英語の原文eが与えられたとき、金の翻訳の長さをmとすると、RCTM IIによって計算された分布を長さmのすべてのセンテンスに対して探索する。RCTM IIの分布から2000個の文を置換しながらサンプリングし、それぞれ1つずつ単語を予測して得られた。まず、最初のターゲット単語の分布を予測し、その分布を最も可能性の高い上位5つの単語に制限して、制限された5つの単語の分布から翻訳候補の最初の単語をサンプリングします。残りの単語についても同様に行います。サンプリングされた各文には，モデルによって明確に定義された確率が割り当てられているため，順位付けを行うことができます。表3は、様々な英語の原文と、RCTM IIによって生成されたいくつかのフランス語の翻訳候補を、その順位とともに示している。

タブ3の結果は、候補となる翻訳の顕著な統語的一致を示しています。3は、候補となる翻訳の顕著な統語的一致を示しています。

英語	原文フランス語金 翻訳ランク	翻訳RCTM II候補	
患者は病気である ...	le patient est malade	...le patient est insuffisante 患者は死亡しました 患者は不健康である	.1 。4 。23
患者は死んでいる ...	le patient est mort	...le patient est mort le patient est dé.	.1 4
患者は病気です ...	le patient est malade	...le patient est mal	.3
患者は病気です。	les patients sont malades .	les patients sont és . 患者さんたちは苦しんでいます	2 。5
患者たちは死んでいる ...	les patients sont morts	... les patients sont morts	...1
The patients are ill .	les patients sont malades .	les patients sont és .	5
その患者は病気でした。	le patient était malade .	le patient était mal .	2
The patients are not dead ...	les patients ne sont pas morts	...les patients ne sont pas morts	.1
患者は病気ではない .	les patients ne sont pas malades	.les patients ne sont pas ()*unknown* 患者さんたちは悪くありません	.1 。6
患者は救われた。	les patients ont été és .	Les patients ont été ées. （患者は救われた	6

表3：英語の原文，フランス語のそれぞれの訳文，RCTM IIから生成した候補訳文で，2000個のサンプルの中から確率の低下に応じてランク付けしたもの。なお，文末のドット（・）は翻訳の一部として生成されたものである．

IIは英語の原文からかなりの量の統語的・意味的情報を取り込み、それをフランス語の翻訳文にうまく移行させることができると考えられる。

表4：単語ペナルティWPで線形補間した各RCTMのWMT-
NTセットでのBleuスコア。cdecシステムには、WPの他に、5つの翻訳モデルと2つの言語モデリング機能などが含まれています。

翻訳候補文の大部分は、完全に整形されたフランス語の文である。さらに、名詞の単数形や複数形、動詞の現在形や過去形などの微妙な構文の特徴は、英語のソースとフランス語のターゲット候補との間でよく相関しています。相関関係が見られない場合や、ターゲットとなる単語がフランス語の語彙に含まれていない場合は、関連する単語や同義語がモデルによって選択されます。これらの特徴から、RCTM

WMT-NT	*2009*	*2010*	*2011*	*2012*
RCTM I + WP	19.7	21.1	22.5	21.5
RCTM II + WP	19.8	21.1	22.5	21.7
cdec (12 features)	19.9	21.2	22.6	21.8

別のシステムによって生成されたものです。本研究では、cdecを用いて、4つのWMT-NTセットの各英語文に対して、1000個の最適な翻訳候補リストを生成した。cdec は、5つの翻訳モデル、2つの言語モデル、1つの単語ペナルティ機能(WP)を含む、12の人工的な特徴を採用しています。RCTMでは、モデルが訳語候補に付与した対数確率を、検証データで調整した単語ペナルティ機能WPで補うだけである。実験の結果は，表4に示すとおりである．4.

結果として得られたBleuのスコアにはほとんどばらつきがありませんでしたが、RCTMのパフォーマンスは、その確率が翻訳の質と相関していることを示しています。単言語のRLM機能とRCTMを組み合わせてもスコアは向上せず、一方で、cdec を1つのコア翻訳確率と言語モデル機能だけに絞ると、スコアは10分の2から5分の1に低下しました。これらの結果は、RCTMが翻訳と言語モデルの両方の分布を学習できたことを示しています。

5.4Rescoring and BLEU評価

第4の実験では、RCTMの能力を検証します。とRCTM IIを比較して、最適な翻訳を選択しています。数多くの翻訳候補の中から

6 結論

我々は、純粋に継続的な文レベルの翻訳モデルのクラスを構成するRecurrent Continuous Transla-tion
Modelsを導入した。これらのモデルの翻訳能力と、参照翻訳に対するパープレクシティの低さを示しました。また、これらのモデルは、構文や意味の情報を取り込み、翻訳候補の品質をリランキング中に推定する能力があることを示した。

RCTMは、連続的な表現が条件付け情報に敏感であるため、非常に柔軟なモデリングが可能である。また、単文を超えた談話表現や、多言語のソース表現を含めることができたり、文字レベルの再帰によって形態学的に豊かな言語をモデル化することができたりするなど、幅広い潜在的な利点や拡張性を示唆しています。

リファレンス

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian
Janvin.2003.ニューラル確率的言語モデル．このように、本稿では、「*機械学習の研究*」をテーマとしています。

ピーター・F・ブラウン、ヴィンセント・J・デラ・ピエトラ、スティーブン・A・デラ・ピエトラ、ロバート.L.
Mercer.1993.統計的機械翻訳の数学。Parameter Estima- tion.*計算言語学*』19:263-311。

R.Collobert and J. Weston.2008.自然言語処理のための統一されたアーキテクチャ。マルチタスク学習による深層ニューラルネットワーク。*国際機械学習会議（ICML）*。

John Duchi, Elad Hazan, and Yoram Singer.2011.オンライン学習と確率的最適化のための適応的劣勾配法。*J. Mach.Learn.Res.* , 12:2121-2159, July.

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitke-vitch, Phil Blunsom, and Philip Resnik.2010. cdec:cdec:
有限状態および文脈自由翻訳モデルのためのデコーダ，アラインメント，および学習フレームワーク．このような状況下では、私たちはこのような問題に対処することができません。これは、計算言語学協会が主催する「ACL2010システム・デモンストレーション」での発表です。

Chris Dyer, Victor Chahuneau, and Noah A. Smith.2013.ibm model 2の簡単、高速、効果的な再パラメータ化。In *Proc. of NAACL*.

Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman.2011.Con-

crete sentence spaces for compositional distributional models of meaning.*CoRR*, abs/1101.0309.

カール・モーリッツ・ヘルマンとフィル・ブランソム。2013.The Role of Syntax in Vector Space Models of Compositional Semantics.この論文は、*計算言語学協会の第51回年次総会の議事録（Vol-ume 1: Long Papers）*、ソフィア、ブルガリア、8月。計算言語学協会.Forthcoming.

Nal KalchbrennerとPhil Blunsom。2013.Recurrent Convolutional Neural Networks for Discourse Com-positionality.談話の位置関係を把握するための再帰的な畳み込みニューラルネットワーク。このようにして、私たちは自分たちの生活をより豊かにすることができるのです。

Hai Son Le, Alexandre Allauzen, and Fran̤cois Yvon.2012.ニューラルネットワークを用いた連続空間翻訳モデル*HLT-NAACL*では、39-48ページ。

Tomas MikolovとGeoffrey Zweig。2012.文脈依存型のリカレントニューラルネットワーク言語モデル。in *SLT*, pages 234-239.

Tomas Mikolov, Martin Kar̈afit, Lukas Burget, Jan Cer- ̈nocky, and Sanjeev Khudanpur. 2010.リカレントニューラルネットワークベースの言語モデル.このように、本稿では、「言語モデル」と「ニューラルネットワーク」の関係について説明します。ISCA.

Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cer- ̈nocky, and Sanjeev Khudanpur. 2011.リカレントニューラルネットワーク言語モデルの拡張。*ICASSP*, pages 5528-5531.IEEE.

Holger Schwenk, Daniel Dchelotte, and Jean-Luc Gauvain.2006.統計的機械翻訳のための連続空間言語モデルin *ACL*.

Holger Schwenk.2012.フレーズベースの統計的機械翻訳のための連続空間翻訳モデル.本論文では、このような問題を解決するための方法を提案する。

Richard Socher, Eric H. Huang, Jeffrey Pennin, Andrew Y. Ng, and Christopher D. Manning.2011.パラフレーズ検出のための動的なプーリングとアンフォールディングの再帰的オートエンコーダ。J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Wein-berger, editors, *Advances in Neural Information Processing Systems 24*, pages 801-809.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng.2012.再帰的な行列-ベクトル空間による意味的構成力。2012年に開催された*Empirical Meth-ods in Natural Language Processing (EMNLP)のPro-ceedings*である。

Ilya Sutskever, James Martens, and Geoffrey E. Hinton.2011.リカレントニューラルネットワークによるテキストの生成Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 1017-1024.Omnipress.