句構造へのアテンションに基づくニューラル機械翻訳モデル

江里口 瑛子　　橋本 和真　　鶴岡 慶雅
東京大学 工学系研究科
{eriguchi, hassy, tsuruoka}@logos.t.u-tokyo.ac.jp

1 はじめに

近年，系列データを固定長ベクトルへ写像し，その固定長ベクトルから異種の系列データへ変換するエンコーダ・デコーダ (Encoder-Decoder; ED) モデルが盛んに研究されている．ニューラルネットワークによる機械翻訳 (Neural Machine Translation; NMT) モデルでは，ニューラルネットワークを利用した ED モデルを用いて，原言語の系列データから目的言語の系列データへの変換を行う．その中でも，エンコーダ側の各隠れ層に注目しながらデコーダの処理を行うアテンションモデルは，機械翻訳タスクの 1 つである WMT'14 の英独翻訳において最高精度を報告している [4]．このような新たな NMT モデルの提案にとどまらず，目的言語のニューラル言語モデルによるデコーダ情報の拡充 [1] や，翻訳以外のタスクと同時学習したマルチタスク学習 [3] などに関する研究が報告されてきている．

一部の NMT モデルは，英独などの近縁の言語対に対して，従来の統計的機械翻訳手法と同等以上の精度を達成している．一方で，英日などの遠縁の言語対に対しては，事前並び替えや構文構造を考慮することで翻訳精度が改善することが知られている [5] が，構文構造を陽に利用した NMT モデルは提案されていない．

本稿では，既存のアテンションモデル [4] を拡張し，エンコーダ側に句構造情報を取り入れる．更に，その句構造へのアテンション機構を取り入れた，新たな NMT モデルの提案を行う．提案手法の特徴は，エンコーダ側に構文情報として句構造を取り入れたニューラル機械翻訳モデルであること，並びに，句へのアテンション機構を有することである．特に，後者により，目的言語側の単語出力と原言語側の句との対応関係を学習することができる．実際，提案モデルを用いて英日翻訳の実験を行ったところ，特に，句構造のみに注目したモデルでは，日本語の単語訳 "手順" を出力する際には "production procedure" という英語の句が，"圧力計 の" に含まれる単語 "の" を出力する際には "of this pressure gage" という句が重要視されており，単語出力と句の対応関係を学習できていることがわかった．

2 系列に注目したニューラル機械翻訳

2.1 エンコーダ・デコーダモデル

ED モデルでは，原言語の文 (単語の系列データ; $\boldsymbol{x} = x_1, x_2, \cdots, x_{T_x}$) を固定長ベクトル空間に埋め込み (**エンコーダ**)，その固定長ベクトル空間から目的言語の文 (単語の系列データ; $\boldsymbol{y} = y_1, y_2, \cdots, y_{T_y}$) を出力する (**デコーダ**)．$j$ 番目の出力単語の条件付き確率は，

$$p(y_j|\boldsymbol{y}_{<j}, \boldsymbol{x}) \quad = \quad g(y_{j-1}, \boldsymbol{s}_j), \tag{1}$$

と計算され，ステップ j でのデコーダの隠れ層 $\boldsymbol{s}_j$ は，

$$\boldsymbol{s}_j \quad = \quad f(y_{j-1}, \boldsymbol{s}_{j-1}), \tag{2}$$

として，1 ステップ前の隠れ層 s_{j-1} と出力単語 y_{j-1} から算出される．ここで，g，f はそれぞれ非線形関数を表す．

隠れ層に Long Short-Term Memory (LSTM) ユニットを採用した ED モデルに，Sequence to sequence (SeqToSeq) モデル [6] がある．ステップ t における入力単語 x_t が n 次元ベクトル (単語ベクトル $\boldsymbol{x}_t$) で表されるとき，LSTM ユニットの隠れ層 $\boldsymbol{h}_t$ は，

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}^{(i)}\boldsymbol{x}_t + \boldsymbol{U}^{(i)}\boldsymbol{h}_{t-1} + \boldsymbol{b}^{(i)}),\\
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1} + \boldsymbol{b}^{(f)}),\\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}^{(o)}\boldsymbol{x}_t + \boldsymbol{U}^{(o)}\boldsymbol{h}_{t-1} + \boldsymbol{b}^{(o)}),\\
\tilde{\boldsymbol{c}}_t &= \tanh(\boldsymbol{W}^{(\tilde{c})}\boldsymbol{x}_t + \boldsymbol{U}^{(\tilde{c})}\boldsymbol{h}_{t-1} + \boldsymbol{b}^{(\tilde{c})}),\\
\boldsymbol{c}_t &= \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1},\\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t),
\end{aligned} \tag{3}$$

と算出される．ここで，$\boldsymbol{i}_t$，$\boldsymbol{f}_t$，$\boldsymbol{o}_t$，$\tilde{\boldsymbol{c}}_t$，$\boldsymbol{c}_t \in \mathbb{R}^{n\times 1}$ は，それぞれ，入力ゲート，忘却ゲート，出力ゲート，更新用メモリセル，メモリセルを表す．$\boldsymbol{W}^{(i)}$，$\boldsymbol{W}^{(f)}$，$\boldsymbol{W}^{(o)}$，$\boldsymbol{W}^{(\tilde{c})} \in \mathbb{R}^{n\times n}$，並びに，$\boldsymbol{U}^{(i)}$，$\boldsymbol{U}^{(f)}$，$\boldsymbol{U}^{(o)}$，$\boldsymbol{U}^{(\tilde{c})} \in \mathbb{R}^{n\times n}$ は重み行列，$\boldsymbol{b}^{(i)}$，$\boldsymbol{b}^{(f)}$，$\boldsymbol{b}^{(o)}$，$\boldsymbol{b}^{(\tilde{c})} \in \mathbb{R}^{n\times 1}$ はバイアス項である．n は単語ベクトルと隠れ層の次元である．σ，tanh，$\odot$ は，それぞれ，ロジスティックシグモイド関数，ハイパボリックタンジェント，ベクトルの要素積を表す．

2.2 系列への注目学習モデル

アテンションに基づくニューラル機械翻訳 (Attention-based Neural Machine Translation; ANMT) モデル [4] は，出力単語を予測する際にエンコーダ側の各隠れ層の寄与分を考慮する，アテンション機構を有している．まず，目的言語における単語の予測を行うデコーダとして，新たに，エンコーダの各隠れ層の注目度合い (アテンション) を考慮した $\tilde{\boldsymbol{s}}$ を用意する．j 番目のデコーダの隠れ層 $\tilde{\boldsymbol{s}}_j$ は，

$$\tilde{s}_j = \tanh(\boldsymbol{W}_c[\boldsymbol{s}_j;\ \boldsymbol{d}_j] + \boldsymbol{b}_c), \tag{4}$$

と算出される．式 (4) の; は，行列同士の結合を表す．この $\tilde{\boldsymbol{s}}_j$ を用いて，j 番目の出力単語の予測分布を，

$$p(y_j|y_{<j}, \boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}_s\tilde{\boldsymbol{s}}_j + \boldsymbol{b}_s), \tag{5}$$

とする．ここで，$\boldsymbol{W}_c \in \mathbb{R}^{n\times 2n}$，$\boldsymbol{W}_s \in \mathbb{R}^{|V|\times n}$ は，重み行列，$\boldsymbol{b}_c \in \mathbb{R}^{n\times 1}$，$\boldsymbol{b}_s \in \mathbb{R}^{|V|\times 1}$ はバイアス項である．$\tilde{\boldsymbol{s}}_j$，$\boldsymbol{d}_j$ の次元はそれぞれ n とし，$|V|$ は目的言語の語彙数を表す．$\boldsymbol{d}_j$ はステップ j における文脈ベクトルであり，

$$\boldsymbol{d}_j \quad = \quad \sum_{i=1}^{T_x} \boldsymbol{\alpha}_j(i)\boldsymbol{h}_i, \tag{6}$$

の形で，アライメントスコア $\boldsymbol{\alpha}_j(i)$ によるエンコーダの各隠れ層 $\boldsymbol{h}_i$ $(i = 1, \cdots, T_x)$ の重み付き和として算出する．ここで，$\boldsymbol{\alpha}_j(i)$ は，ステップ i におけるエンコーダの隠れ層 $\boldsymbol{h}_i$ とステップ j における SeqToSeq モデルのデコーダ $\boldsymbol{s}_j$ とのアライメントスコアを表し，

$$\boldsymbol{\alpha}_j(i) \quad = \quad \frac{\exp(\boldsymbol{h}_i \cdot \boldsymbol{s}_j)}{\sum_k^{T_x} \exp(\boldsymbol{h}_k \cdot \boldsymbol{s}_j)}, \tag{7}$$

として，最終的にソフトマックス関数を用いて全体の和が 1 となるよう正規化された確率分布の形とする．$\boldsymbol{h}_i \cdot \boldsymbol{s}_j$ は，エンコーダ並びにデコーダの隠れ層の類似度を表す．

2.3 目的関数とパラメータの学習方法

ANMT モデルの目的関数は，学習データ $\mathcal{D}$ の入力文と出力文の全ペアに関する対数尤度として，

$$J(\boldsymbol{\theta}) \quad = \quad \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \log p(\boldsymbol{y}|\boldsymbol{x}), \tag{8}$$

とする．ここで $\boldsymbol{\theta}$ は，モデルが含む LSTM，単語ベクトル，アテンションなどの計算に用いられる全てのパラメータを表す．学習時は，パラメータ更新に確率的勾配降下法を用いる．

3 句構造に注目したニューラル機械翻訳

語順の似通った英仏翻訳に対して ED モデルを適用した際，原言語の語順を逆順に入力することで翻訳精度向上に繋がったという報告がある [6]．日本語と英語のように語順や文法構造が大きく異なる言語対は多く存在し，系列データのみを扱う従来の NMT モデルによって学習を行うことは難しいと考えられる．例えば，日本語の語彙には，英語の定冠詞 (“the”) に一対一対応する明示的な単語はない．しかしながら，句構造の観点から，日本語の “月”と英語の “the moon”は名詞句としてアライメントをとることができる．

ANMT モデルにおけるアテンション機構は，デコーダの出力した単語とエンコーダ側の隠れ層とのアライメントの役割を担う．単語が入力されたエンコーダ側の隠れ層は，デコーダの隠れ層に対して各々独立にアライメントスコアが算出されており，句のような構文構成要素の観点からのアライメント学習は行われていない．Luong ら [4] は，デコード時に注目する領域をある窓幅 $|n|$ 内の系列データ (n-gram) に限定した，局所的 ANMT モデルの提案を行っているが，最適窓幅 $|n|$ は実験的に一意に求める必要があり，また，その窓幅外の情報は一切考慮できないという欠点がある．

我々は，句構造に注目したニューラル機械翻訳モデルを提案する．まず，異なる言語対に対する有効な情報として，新たに，原言語側の句構造をデコーダへの情報として加える．さらに，単語 (葉ノード) からボトムアップに構成されていく句 (ノード) へのアテンション機構を導入する．これにより，先行研究で行われていた出力単語と入力単語の単語対応関係学習に加えて，出力単語と原言語側の句の対応関係の学習もまた同時に行うことができる．図 1 に提案モデルの概要図を示す．

3.1 句構造情報のベクトル表現

2 分木構造で表された句構造情報に従い，文を構成している句 (ノード) の隠れ層を算出する．句の隠れ層には Tree Long Short-Term Memory (Tree LSTM) ユニット [7] を用いる．N 個の LSTM ユニットを子ノードに持つとき，ノード番号 j の親ノードの隠れ層 $\boldsymbol{h}_j$ は，

$$\begin{aligned}
\boldsymbol{i}_j &= \sigma(\boldsymbol{W}^{(i)}x_j + \sum_{l=1}^{N} \boldsymbol{U}_l^{(i)}\boldsymbol{h}_{jl} + \boldsymbol{b}^{(i)}),\\
\boldsymbol{f}_{jk} &= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_j + \sum_{l=1}^{N} \boldsymbol{U}_{kl}^{(i)}\boldsymbol{h}_{jl} + \boldsymbol{b}^{(f)}),\\
\boldsymbol{o}_j &= \sigma(\boldsymbol{W}^{(o)}\boldsymbol{x}_j + \sum_{l=1}^{N} \boldsymbol{U}_l^{(o)}\boldsymbol{h}_{jl} + \boldsymbol{b}^{(o)}),
\end{aligned}$$

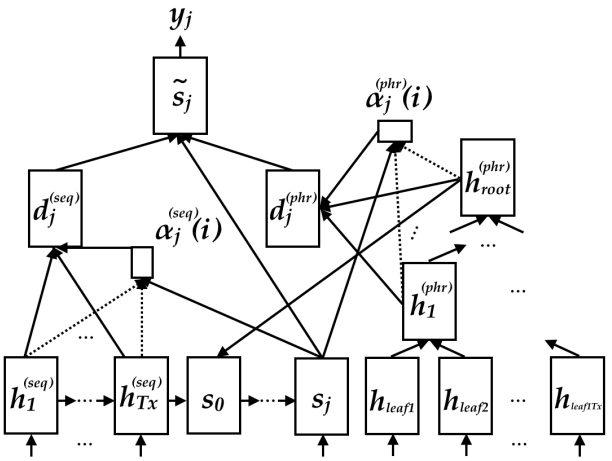

図 1: 提案モデル.

$$\begin{aligned}\tilde{\boldsymbol{c}}_j &= \tanh(\boldsymbol{W}^{(\tilde{c})}\boldsymbol{x}_j + \sum_{l=1}^{N}\boldsymbol{U}_l^{(\tilde{c})}\boldsymbol{h}_{jl} + \boldsymbol{b}^{(\tilde{c})}),\\ \boldsymbol{c}_j &= \boldsymbol{i}_j \odot \tilde{\boldsymbol{c}}_j + \sum_{l=1}^{N}\boldsymbol{f}_{jl}\odot\boldsymbol{c}_{jl},\\ \boldsymbol{h}_j &= \boldsymbol{o}_j \odot \tanh(\boldsymbol{c}_j), \end{aligned} \quad (9)$$

と算出される．ここで，$\boldsymbol{i}_j$，$\boldsymbol{f}_{jk}$，$\boldsymbol{o}_j$，$\tilde{\boldsymbol{c}}_j$，$\boldsymbol{c}_j$，$\boldsymbol{x}_j \in \mathbb{R}^{n\times 1}$，並びは，それぞれ，入力ゲート，ノード番号 l $(1 \le l \le N)$ の子ノード用忘却ゲート，出力ゲート，更新用メモリセル，メモリセル，単語ベクトルを表す．$\boldsymbol{W}^{(i)}$，$\boldsymbol{W}^{(f)}$，$\boldsymbol{W}^{(o)}$，$\boldsymbol{W}^{(u)} \in \mathbb{R}^{n\times n}$，並びに，$\boldsymbol{U}_l^{(i)}$，$\boldsymbol{U}_l^{(f)}$，$\boldsymbol{U}_l^{(o)}$，$\boldsymbol{U}_l^{(u)} \in \mathbb{R}^{n\times n}$ は，各ゲート及び更新用メモリセルにおける重み行列，$\boldsymbol{b}^{(i)}$，$\boldsymbol{b}^{(f)}$，$\boldsymbol{b}^{(o)}$，$\boldsymbol{b}^{(u)} \in \mathbb{R}^{n\times 1}$ は，バイアス項である．n は単語ベクトルと隠れ層の次元である．句構造情報は 2 分木で与えられるため，$N = 2$ である．また，初めて単語を隠れ層へ埋め込む際の初期状態は $\boldsymbol{h}_{jl} = \boldsymbol{0}$ とし，句の隠れ層への遷移では $\boldsymbol{x}_j = \boldsymbol{0}$ として算出する．

3.2 系列と句構造への同時注目学習モデル

j 番目のデコーダ $\tilde{\boldsymbol{s}}_j$ に対する，系列データ (seq) の文脈ベクトル $\boldsymbol{d}_j^{(seq)}$ は系列データの隠れ層 $\boldsymbol{h}_i^{(seq)}$ から，

$$\boldsymbol{d}_j^{(seq)} = \sum_{i=1}^{T_x^{(seq)}} \boldsymbol{\alpha}_j^{(seq)}(i)\boldsymbol{h}_i^{(seq)}, \quad (10)$$

と算出される．ここで，$T_x^{(seq)}$ は隠れ層 $\boldsymbol{h}^{(seq)}$ の総数である．同様に，句構造 (phr) の文脈ベクトル $\boldsymbol{d}_j^{(phr)}$ は句構造の隠れ層 $\boldsymbol{h}_i^{(phr)}$ から，

$$\boldsymbol{d}_j^{(phr)} = \sum_{i=1}^{T_x^{(phr)}} \boldsymbol{\alpha}_j^{(phr)}(i)\boldsymbol{h}_i^{(phr)}, \quad (11)$$

と算出される．ここで，$T_x^{(phr)}$ $(= T_x^{(seq)} - 1)$ は句の隠れ層 $\boldsymbol{h}^{(phr)}$ の総数である．最後に，デコーダ $\tilde{\boldsymbol{s}}_j$ を，

$$\tilde{\boldsymbol{s}}_j = \tanh(\boldsymbol{W}_e[\boldsymbol{s}_j;\ \boldsymbol{d}_j^{(seq)};\ \boldsymbol{d}_j^{(phr)}] + \boldsymbol{b}_e), \quad (12)$$

と算出する．$\boldsymbol{W}_e \in \mathbb{R}^{n\times 3n}$ は重み行列，$\boldsymbol{b}_e \in \mathbb{R}^{n\times 1}$ はバイアス項である．提案モデルでは，2 種類の文脈ベクトル $\boldsymbol{d}_j^{(seq)}$，$\boldsymbol{d}_j^{(phr)}$ をデコーダ $\tilde{\boldsymbol{s}}_j$ への入力として与え，この際，両者の重要性は等価であるとしている．

3.3 デコーダの初期状態設定

デコーダの初期状態 $\boldsymbol{s}_0$ は，Tree LSTM を用いて，系列データから算出される最終 LSTM ユニット $\boldsymbol{h}_{T_x}^{(seq)}$ と，句構造情報から算出される最終 LSTM ユニット $\boldsymbol{h}_{root}^{(phr)}$ (ルートの隠れ層) を子ノードとする親ノードとして求める．ここで用いた Tree LSTM ユニットは 3.1 節のユニットとは別に，新たに用意したものである．

Tree LSTM は，N 個の LSTM ユニット (子ノード) が親ノードの LSTM ユニットへ合流されて構成される．このため，LSTM ユニットを利用したアテンション機構の拡張を行う場合，句構造への同時注目学習で示したように，Tree LSTM を用いて，それらの全体情報をデコーダの初期状態へ入力として追加することは容易である．

4 実験

4.1 コーパス

使用したコーパスは，Asian Scientific Paper Excerpt Corpus (ASPEC)[1] の英日コーパスである．文の句構造は，HPSG 文法に基づく解析器 Enju [2] を用いて，構文解析が成功した文から抽出した．それ以外の場合は，句構造情報からデコーダの初期状態への入力は $\boldsymbol{h}_{root}^{(phr)} = \boldsymbol{0}$ であり，j 番目の句構造の文脈ベクトルは $\boldsymbol{d}_j^{(phr)} = \boldsymbol{0}$ である．英語の単語分割は Enju の解析結果を用い，日本語の単語分割は MeCab [3] を用いた．原言語，並びに目的言語の学習コーパスから，1 文あたり 20 単語を超える文対を削除したところ，学習データ (train-1.txt; 100 万文対) から 232,086 文対 (このうち，構文解析が成功した英文は 230, 778 文) を抽出した．語彙は，学習データ中の 94, 744 語 (英語), 60, 119 語 (日本語) の全てを用いた．また，開発データ (dev.txt; 1790 文対) のうち解析が成功した英文は 1,779 文であった．

[1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[2] http://kmcs.nii.ac.jp/enju/
[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

4.2 モデルパラメータの学習

モデル中のパラメータは，$[-0.1, 0.1]$ を範囲とする一様分布からの乱数に従って初期化を行い，バイアス項は $\mathbf{0}$ とした．各パラメータの学習は，確率的勾配降下法 (学習率は 0.5) を用い，ミニバッチサイズを 50 とした．Sutskever ら [6] に倣い，勾配ノルムは 5 でクリップした．また，単語ベクトル，隠れ層の次元は全て $n = 200$ とした．学習の高速化のため，負例サンプリングに基づくソフトマックスの近似手法である BlackOut [2] を用いた．学習後，BlackOut は通常のソフトマックス関数として用いることができ，RNN 言語モデルの学習においてその有効性が示されている．BlackOut のパラメータは，$K = 200$，$\alpha = 0.4$ とした．

5 結果・考察

英語から日本語への翻訳 (英日翻訳) の実験を，(a) 系列に注目した ANMT モデル，(b) 句構造のみに注目したモデル (提案手法)，(c) 系列と句構造に同時注目したモデル (提案手法) の 3 手法をそれぞれ用いて行った．ただし，系列データを利用した手法 (a)，(c) では，単語の入力順を順方向と，逆順方向 (逆順) の各場合で実験を行った．手法 (b) では，アテンションの対象範囲に単語 (葉ノード) も含めた．表 1 に，開発データ (解析成功文) の参照訳に対する 1 文あたりの最大対数尤度をモデル別にまとめる．この表より，句構造をモデルに取り入れることの有効性が示された．

図 2 は，全モデル中で最大の対数尤度を示した手法 (b) を用いて，構文解析が成功した英文の開発データを与えた際の，日本語の翻訳結果例と，句への重み付きアテンションの例である．翻訳の出力はビーム探索により行い，ビーム幅は 50 とした．日本語の名詞 “手順”は，英語の句 “production procedure”と $\alpha^{(phr)} = 0.53$ の重みで，また，動詞 “述べ”は，英語の単語 “describes”に $\alpha^{(phr)} = 0.33$ の重みで，それぞれ対応付けられている．この他のアテンションの例として，“圧力 計 の”の “の”へ対応する句に “of this pressure gage” ($\alpha^{(phr)} = 0.17$) があり，出力単語と句の対応関係が句構造へのアテンション機構により学習されていることがわかる．

6 おわりに

本研究では，句構造へ注目した新たな ANMT モデルを提案した．さらに，英日翻訳の実験を通して，既存の ANMT 手法と比較し，句構造へのアテンションモデルの有効性を示した．提案モデルにおける Tree LSTM の葉ノード (単語) には，周辺文脈情報が含まれていないため，その不足を補填する必要がある．今後は，より大規模なデータを用いた実験を行い，提案モデルの定量的評価を翻訳指標 BLEU などによって行う．

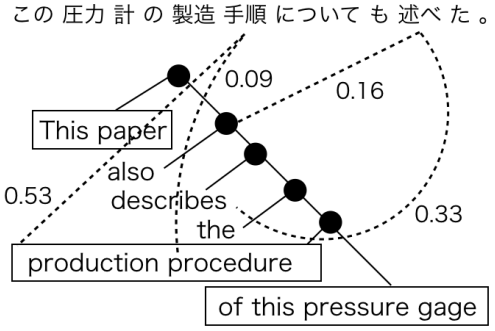

図 2: 開発データの翻訳例とアテンションの様子.

モデル	平均対数尤度
(a-1) ANMT [4]	-117.8
(a-2) ANMT [4] (逆順)	-121.7
(b) **提案手法** (句)	**-109.7**
(c-1) **提案手法** (系列+句)	-115.0
(c-2) **提案手法** (系列+句; 逆順)	-116.0

表 1: 開発データ 1 文あたりの最大対数尤度.

謝辞

本研究は JSPS 科研費 15J12597 の助成，JST，CREST の支援を受けたものです.

参考文献

[1] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015.

[2] S. Ji, S. V. N. Vishwanathan, N. Satish, M. J. Anderson, and P. Dubey. Blackout: Speeding up recurrent neural network language models with very large vocabularies. *arXiv preprint arXiv:1511.06909*, 2015.

[3] T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015.

[4] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the EMNLP*, 2015.

[5] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proceedings of the ACL*, 2013.

[6] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in NIPS 27*. 2014.

[7] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the ACL and the IJCNLP*, 2015.