学修番号 18860627

修士論文

目的言語側の言語モデルを用いたニューラル機械翻訳

黒澤 道希

2020 年 2 月 21 日

首都大学東京大学院
システムデザイン研究科 情報科学域

黒澤 道希

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　（副指導教員）

高間 康史 教授　（副指導教員）

目的言語側の言語モデルを用いたニューラル機械翻訳*

黒澤 道希

内容要旨

近年機械翻訳の研究において，流暢性の高い出力を得られるニューラル機械翻訳（Neural Machine Translation: NMT）が盛んに研究されている．NMT は原言語文を中間表現に変換する機構（Encoder）と中間表現から目的言語文を生成する機構（Decoder）の 2 つからなるモデルが基本である．NMT が研究され始めた当初は回帰型ニューラルネットワーク（Recurrent Neural Network: RNN）のみによって構成されており，旧来の統計的機械翻訳（Statistical Machine Translation: SMT）と比較して流暢性が飛躍的に向上した．その後，Decoder の出力時に Encoder の情報を注視する注意機構（Attention 機構）の登場により原言語文の情報をより保持した翻訳を出力できるようになった．しかしながら NMT には重複した出力を生成するなど多くの問題も存在しており，さらなる研究が行われている．

その一つに言語モデルを用いた研究がある．言語モデルはその言語らしい出力をする機構であり，単言語で構成されるため比較的流暢な出力を得ることができる．ニューラル言語モデルの構造は NMT の Encoder や Decoder と同じ RNN が一般的であり，構造が等しいため NMT も言語モデルとしての働きを持っていると考えることもできる．一方で，ニューラル言語モデルを単体として用意することで，単一言語に特化した出力を予測することができるようになり，翻訳モデルと合わせて利用することにより，より出力言語らしい流暢な出力が可能となる．

言語モデルを用いる先行研究では，翻訳機構と言語モデル機構の 2 つを用意し双方の情報を用いる．双方の機構の予測を同尺度もしくは動的に重み付けして出力単語を予測することで，言語モデルの情報を翻訳に混ぜ合わせるものや，言語モデルの予測時点から翻訳機構の情報を与えて混ぜ合わせることによって出力単語を予測するものなどがある．しかしながら，機械翻訳においては流暢性の向上だけではな

*首都大学東京大学院 システムデザイン研究科 情報科学域 修士論文, 学修番号 18860627, 2020 年 2 月 21 日.

く妥当性を担保することも求められるため，翻訳機構と言語モデル機構の予測を単純に混ぜ合わせるべきでなく，翻訳機構の情報を活用し言語モデル機構の情報を補助的に用いるべきであるが，先行研究のほとんどは双方の情報を手動で決めた重みを用いるなど，単純な方法で混ぜ合わせて出力単語を予測している．

本研究では，翻訳機構と言語モデル機構の 2 つを用意した上で，翻訳機構を主軸とし言語モデル機構を補助的に活用するモデル: Dynamic Fusion を提案する．本モデルでは言語モデル機構の独立的予測のため言語モデルの予測には翻訳機構の情報は与えない．また，翻訳機構の情報を元にするため注意機構的に言語モデルを活用する．まず，翻訳機構の隠れ層から各単語に対して単語 Attention を用い各単語の重要度を求める．その重要度を言語モデル機構の予測確率を用いて重み付ける．最終的に重み付けされた重要度を元に Attention を利用してモデル全体としての出力を決定する．これにより，翻訳機構の情報はそのまま保持した上で言語モデル機構の情報を活用することができ，妥当性を担保した上で流暢な出力が可能になることが期待される．加えて，先行研究においては語彙同士の予測確率を掛け合わせる手法もありそれぞれの語彙が一致している必要性があったが，提案手法においては Attention によって利用するにとどまっており，単語埋め込みが取得できる限り翻訳機構と言語モデル機構の語彙が必ずしも揃っている必要がない．

提案したモデル及び先行研究のモデルについて日英言語対に対して実験を行い，実験した全ての設定において提案手法が有用であることを示す．加えて，先行研究では不可能であった実用に近い設定においても，言語モデルを用いない機械翻訳と比較して有用であることを示す．本論文ではその自動評価結果を示すとともに実際の出力例を元に分析を行った結果について示す．また，言語モデル機構の予測に対する Attention と実際の出力を分析することにより，言語モデルが文法的性質を用いて予測を補助するために有用な情報である可能性が高いことを合わせて示す．

本論文の主な貢献は以下の通りである．

- 本論文では言語モデルを用いた NMT: Dynamic Fusion を新たに提案した．提案手法は Attention を用いて言語モデルの予測確率を組み合わせるものである．
- 英語–日本語における双方向の翻訳に言語モデルを用いることで流暢かつ妥当な出力が可能であることを示した．

- Dynamic Fusion がより現実に近い設定においても有意に翻訳精度が向上することを示した．
- Dynamic Fusion に関して Attention の重みを中心に翻訳向上に寄与する要因について分析を行った．

本論文の構成は次の通りである．第 1 章では本研究の背景，提案，貢献について述べる．第 2 章ではニューラル機械翻訳の基本的な構造について述べる．第 3 章では言語モデルを用いたニューラル機械翻訳に関する先行研究について述べる．第 4 章では言語モデルを注意機構的に用いたニューラル機械翻訳の手法について述べる．第 5 章では第 3，4 章で述べた手法を用いた実験について述べる．第 6 章では実験の結果およびその考察について述べる．最後に，第 7 章で本研究のまとめについて述べる．

Neural Machine Translation using Language Model of Target Language*

Michiki Kurosawa

Abstract

In recent years, in machine translation research, neural machine translation (NMT), which can obtain highly fluent output, has been actively studied. NMT is based on a model consisting of two mechanisms: a mechanism for converting source language sentences into hidden representations (Encoder) and a mechanism for generating target language sentences from hidden representations (Decoder). When NMT was first studied, it was composed of a recurrent neural network (RNN), and its fluency was dramatically improved compared to the conventional statistical machine translation (SMT). After that, with the advent of an attention mechanism (Attention mechanism) that gazes at the information of the Encoder when outputting the Decoder, it became possible to output a translation with more information on the source language sentence. However, NMT has many problems, such as generating duplicate outputs, and further research is being conducted.

As such, language models have been investigated for incorporation with NMT. In prior investigations, two models have been used: a translation model and a language model. The translation model' s predictions are weighted by a language model with a hand-crafted scale in advance. However, these approaches fail to adopt the language model weighting with regard to the translation history. In another line of approach, language model prediction is incorporated into the translation model by jointly considering source and target informa-

*Master's Thesis, Department of Computer Science, Graduate School of Systems Design. Tokyo Metropolitan University, Student ID 18860627, February 21, 2020.

tion. However, this is limited because it largely ignores the adequacy of the translation output.

In this paper, we propose a "Dynamic Fusion" mechanism that predicts output words by attending to the language model. We hypothesize that each model should make predictions according to only the information available to the model itself; the information available to the translation model should not be referenced before prediction. In the proposed mechanism, a translation model is fused with a language model through the incorporation of word-prediction probability according to the attention. However, the models retain predictions independent of one another. As a result, it is expected that the information of the language model mechanism can be used while retaining the information of the translation mechanism as it is, and fluent output will be possible while ensuring the adequacy. In addition, in previous research, there was a method of multiplying the predicted probabilities of the vocabularies, and it was necessary that each vocabulary be the same. In the proposed method, it is used only by Attention, and the vocabulary of the translation mechanism and the language model mechanism does not necessarily need to be the same as long as word embedding can be obtained.

Experiments are performed on Japanese–English language pairs leveraging the proposed model and the model of the previous research , and it is shown that the proposed method is effective in all settings. In addition, we show that even in practical settings, which were impossible in previous studies, our approach are more useful than machine translation without language models. In this paper, we show the results of the automatic evaluation and the results of analysis based on actual output examples. We also show that the language model is likely to be effective information to assist the prediction by analyzing the Attention for the prediction of the language model mechanism and the output sentences of the model.

The main contributions of this paper are as follow:

- We propose an attentional language model: Dynamic Fusion that effec-

tively introduces a language model to NMT.

- We show that fluent and adequate output can be achieved with a language model in English–Japanese translation.
- We show that Dynamic Fusion significantly improves translation accuracy in a realistic setting.
- Dynamic Fusion' s ability to improve translation is analyzed with respect to the weight of the attention.

The structure of this paper is as follows. Chapter 1 describes the background, proposals, and contributions of this research. Chapter 2 describes the basic structure of neural machine translation. Chapter 3 describes previous research on neural machine translation using language models. Chapter 4 describes a method of neural machine translation using a language model as an attention mechanism. Chapter 5 describes experiments using the methods described in Chapters 3 and 4. Chapter 6 describes the experimental results and their considerations. Finally, Chapter 7 gives a summary of this study.

目次

第 1 章　はじめに

1.1　ニューラル機械翻訳の発展

近年，機械翻訳の研究において，流暢性の高い出力を得られるニューラル機械翻訳（Neural Machine Translation: NMT）が盛んに研究されている．NMT は原言語文を中間表現に変換する機構（Encoder）と中間表現から目的言語文を生成する機構（Decoder）の 2 つからなるモデルが基本である [1, 2]．NMT では回帰型ニューラルネットワーク（Recurrent Neural Network: RNN）によって構成されており，旧来の統計的機械翻訳（Statistical Machine Translation: SMT）と比較して流暢性が飛躍的に向上した [3]．その後，Decoder の出力時に Encoder の情報を注視する注意機構（Attention 機構）の登場により原言語文の情報をより保持した翻訳を出力できるようになった [4, 5]．

1.2　ニューラル機械翻訳における問題点の改善

NMT には重複した出力を生成するなど多くの問題が存在する．中でも NMT の学習には多くの対訳コーパスが必要であり，良質なコーパスの生成にはコストがかかるため，この問題を解決する手法として単言語コーパスを活用する研究が行われている．単言語コーパスは比較的容易に収集が可能であり，また，SMT にも利用されている [6]．単言語コーパスを用いる研究は複数のアプローチが行われており，翻訳機構の事前学習 [7]，単語分散表現の初期化 [8, 9]，逆翻訳による疑似対訳コーパスの生成 [10] などが存在する．本論文では言語モデルを用いたアプローチ [11, 12] に着目する．

1.3　言語モデルによる流暢性の改善

NMT の出力文は流暢である傾向にあるが，翻訳を実行する際には入力文の情報のみが与えられるため，出力言語の言語特性を考慮することは難しい [13]．

そこで言語モデルに着目する．言語モデルはその言語らしい出力をする機構であ

り，単言語で構成されるため比較的流暢な出力を得ることができる [14]．ニューラル言語モデルの構造は NMT の Encoder や Decoder と同じ機構が使われることが多く，構造が等しく NMT も言語モデルとしての働きを持つことが可能であり，新たな言語モデル機構の追加は不要であると考えることもできる．しかしながら，ニューラル言語モデルを単体として用意することで，単一言語に特化した出力を予測することができるようになり，翻訳モデルと合わせて利用することにより，翻訳の情報だけでなく言語モデルによる出力言語の情報も考慮することができるようになり，より出力言語らしい流暢な出力が可能となる．

言語モデルを用いる先行研究では，翻訳機構と言語モデル機構の 2 つを用意し双方の情報を用いる．双方の機構の予測を同尺度もしくは動的に重み付けして出力単語を予測することで言語モデルの情報を翻訳に混ぜ合わせるものや，言語モデルの予測時点から翻訳機構の情報を与えて混ぜ合わせることで出力単語を予測するものなどがある．Shallow Fusion [11] では，双方の機構の予測確率を人手で固定した重みで足し合わせた．Deep Fusion [11] では，双方の機構の隠れ層を，言語モデル機構を元にした重みで足し合わせることによって混ぜ合わせた．Cold Fusion [15] では，言語モデル機構の予測を別の機構で処理を加えた後に，動的に決めた重みで組み合わせた．Simple Fusion [12] では，双方の出力を重みなしに組み合わせた．しかしながら，機械翻訳においては流暢性の向上だけではなく妥当性を担保することも求められるため，翻訳機構と言語モデル機構の予測を単純に混ぜ合わせるべきでなく，翻訳機構の情報を活用し言語モデル機構の情報を補助的に用いるべきであるが，先行研究においては双方の情報を混ぜ合わせて出力単語を予測している．

1.4 本論文の貢献

本研究では，翻訳機構と言語モデル機構の 2 つを用意した上で，翻訳機構を主軸とし言語モデル機構を補助的に活用する **Dynamic Fusion** を提案する．本モデルでは言語モデル機構の独立的予測のため言語モデルの予測には翻訳機構の情報は与えない．また，翻訳機構の情報を元にするため注意機構的に言語モデルを活用する．まず，翻訳機構の隠れ層から各単語に対して単語 Attention を用い各単語の重要度を求める．その重要度を言語モデル機構の予測確率を用いて重み付ける．最終

的に重み付けされた重要度を元に Attention を利用してモデル全体としての出力を決定する．これにより，翻訳機構の情報はそのまま保持した上で言語モデル機構の情報を活用することができ，妥当性を担保した上で流暢な出力が可能になることが期待される．加えて，先行研究においては語彙同士の予測確率を掛け合わせる手法もありそれぞれの語彙が一致している必要性があったが，提案手法においては Attention によって利用するにとどまっており，単語埋め込みが取得できる限り翻訳機構と言語モデル機構の語彙が必ずしも揃っている必要がない．

提案したモデル及び先行研究のモデルについて日英言語対に対して実験を行い，実験した全ての設定において提案手法が有用であることを示す．加えて，先行研究では不可能であった実用に近い設定においても，言語モデルを用いない機械翻訳と比較して有用であることを示す．本論文ではその自動評価結果を示すとともに実際の出力例を元に分析を行った結果について示す．また，言語モデル機構の予測に対する Attention と実際の出力を分析することにより，言語モデルが文法的性質を用いて予測を補助するために有用な情報である可能性が高いことを合わせて示す．

本論文の主な貢献は以下の通りである．

- 本論文では言語モデルを用いた Dynamic Fusion を新たに提案した．提案手法は Attention を用いて言語モデルの予測確率を組み合わせるものである．
- 英語–日本語における双方向の翻訳に言語モデルを用いることで流暢かつ妥当な出力が可能であることを示した．
- Dynamic Fusion がより実用的な設定においても有意に翻訳精度が向上することを示した．
- Dynamic Fusion に関して Attention の重みを中心に翻訳向上に寄与する要因について分析を行った．

本論文の構成は次の通りである．第 2 章ではニューラル機械翻訳の基本的な構造について述べる．第 3 章では言語モデルを用いたニューラル機械翻訳に関する先行研究について述べる．第 4 章では言語モデルを注意機構的に用いたニューラル機械翻訳の手法について述べる．第 5 章では第 3，4 章で述べた手法を用いた実験について述べる．第 6 章では実験の結果およびその考察について述べる．最後に，第 7 章で本研究のまとめについて述べる．

第 2 章　ニューラル機械翻訳

2.1　Encoder–Decoder モデル

Sutskever ら [1] は Long Short–Term Memory（LSTM）を用いた Sequence–to–Sequence のモデルを提案した．その後，Bahdanau ら [4] と Luong ら [5] は，Attention を用いた NMT を提案した．本節では Luong らのモデルを中心に構成したベースラインシステムについて説明する．

このモデルは原言語文を扱う Encoder と目的言語文を出力する Decoder に大別される．Encoder では双方向 LSTM を，Decoder では単方向 LSTM を用いている．

Encoder では原言語文を隠れ層へと変換する．原言語文は各ステップ i 毎に処理され，ステップ i の埋め込み層 e_i^{enc} は，

$$e_i^{enc} = \tanh(W_{\mathrm{x}} x_i) \tag{2.1.1}$$

と表される．ここで W_{x}（$V \times |h|$）は単語を埋め込み層へと変換する重み行列を表し，x_i はステップ i の単語を表す one–hot ベクトルである．

次に順方向の隠れ層 $\overrightarrow{h}$，逆方向の隠れ層 $\overleftarrow{h}$ とそれらを合わせた各ステップの隠れ層 $\bar{h}$ はそれぞれ，

$$\overrightarrow{h_i} = \mathrm{LSTM}(e_i^{enc}, \overrightarrow{h_{i-1}}) \tag{2.1.2}$$
$$\overleftarrow{h_i} = \mathrm{LSTM}(e_i^{enc}, \overleftarrow{h_{i-1}}) \tag{2.1.3}$$
$$\bar{h_i} = \overrightarrow{h_i} + \overleftarrow{h_i} \tag{2.1.4}$$

Decoder では Encoder で計算された隠れ層を利用して出力文を生成する．Decoder における隠れ層 h_j は，

$$h_1 = \overrightarrow{h_N} + \overleftarrow{h_1} \tag{2.1.5}$$
$$h_j = \mathrm{LSTM}([e_{i-1}^{dec}; \tilde{h}_{j-1}], h_{j-1})(j \neq 1) \tag{2.1.6}$$

と表される．ここで e_{i-1}^{dec} は 1 ステップ前に出力された単語埋め込み表現を表し，$\tilde{h}_{j-1}$ は 1 ステップ前の Attention を考慮した隠れ層を表す．また，$[a; b]$ は a と b のベクトルの結合を表す．

Decoder の単語埋め込み表現 e_i^{dec} は，

$$e_i^{dec} = \tanh(W_{\mathrm{y}} \hat{y_i}) \tag{2.1.7}$$

と表される．ここで W_{y}（$|V| \times |h|$）は単語を埋め込み層へと変換する重み行列を表し，$|V|$ と $|h|$ はそれぞれ単語数と隠れ層の次元数を表す．また，$\hat{y_i}$ はステップ i の予測単語を表し，学習時には正しい単語を予測できたとして次ステップの計算を行うため，$\hat{y_i}$ の代わりに正解単語 y_i を用いる．

Attention を考慮した隠れ層 $\tilde{h}_j$ は，

$$\tilde{h}_j = \tanh(W_a[h_j; c_j]) \tag{2.1.8}$$

と表される．ここで，W_a（$2|h| \times |h|$）はニューラルネットワークの重みを表し，c_j は Attention の隠れ層を表す．なお，Attention の計算方法については次節で説明する．

最終的な予測単語 $\hat{y_i}$ は，

$$\hat{y_i} = \underset{y}{\operatorname{argmax}} \operatorname{softmax}(W_d \tilde{h}_j) \tag{2.1.9}$$

と表される．ここで W_d（$|h| \times |V|$）はニューラルネットの重みを表す．

2.2 Attention 機構

Bahdanau ら [4] と Luong ら [5] は複数の Attention 機構を提案した．この機構は長文の翻訳において入力側の情報が希釈され忘れられることを避けるために導入されたもので，出力時に必要な入力文の単語を注視する機構である．そのため，現時点での隠れ層を元に必要な（入力）部分を取り出すことができる機構と考えることができる．本論文では Luong らが提案した dot Attention を用いる．

Attention 機構は Encoder の各隠れ層の重み付き和を用いるものであり，Attention の隠れ層 c_j は，

$$c_j = \sum_{i=1}^{|X|} \alpha_{ij} \bar{h_i} \tag{2.2.1}$$

と表される．ここで，α_{ij} は各隠れ層の重みを表し，ソフトマックス関数を用いて和が 1 になるように正規化されており，

$$\alpha_{ij} = \frac{\exp(\bar{h_i}^T h_j)}{\sum_{k=1}^{|X|} \exp(\bar{h_k}^T h_j)} \tag{2.2.2}$$

と表される．

2.3 評価

翻訳の評価については人手による評価と自動評価の 2 つに大別される．

人手による評価は流暢性と妥当性を元に評価されることが多い．流暢性は出力文のみを評価対象に出力言語としての流暢性を評価する．妥当性は入力文の内容が出力文に反映されているかを評価する．しかし，人手による評価にはコストが発生する．

自動評価は主に参照訳との比較で自動的に評価するため，低コストで評価することが可能である．本論文では 2 つの自動評価尺度を用いている．

1 つ目は BLEU [16] である．この尺度は参照訳と出力文の n–gram 一致率を元に評価する尺度である．BLEU は，

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{2.3.1}$$

で計算される．ここで p_n は n–gram 精度[1]を表しており，w_n は各 n–gram の重みを表している．一般的に $N = 4, w_n = \frac{1}{N}$ に設定されており，4–gram までの幾何平均を計算している．また，BP は短い出力文へのペナルティを表し，

$$\mathrm{BP} = \begin{cases} 1 & (c > r) \\ e^{(1-r/c)} & (c \leq r) \end{cases} \tag{2.3.2}$$

で計算される．ここで c は出力文の文長を表し，r は参照訳の文長を表す．この評価尺度には表層の情報のみを考慮できないという欠点もあるが，機械翻訳におけるほとんどの論文で利用されている．

[1]参照訳中に出現した単語は 1 度しか使われない修正 n–gram を使用する．

2 つ目は Rank–based Intuitive Bilingual Evaluation Score (RIBES) [17] である．この尺度は，参照訳と出力文に共通して出現する単語の順序を順位相関係数で評価する尺度である．RIBES は，

$$\text{RIBES} = \text{NSR} \times P^{\alpha} \tag{2.3.3}$$

で計算される．ここで，NSR はスピアマンの順位相関係数を $[0, 1]$ の値に正規化したものである．また，共通単語が少ない場合のペナルティとして P^{α} が存在し，

$$P^{\alpha} = n/h \times \alpha \tag{2.3.4}$$

で表される．ここで n は出力文の中で参照訳にも出現する単語数を表し，h は出力文の単語数を表す．また，α は重みを表すパラメータで一般的に $\alpha = 0.25$ が用いられる．

第 3 章　先行研究

3.1　Shallow Fusion

Gulcehre ら [11] は，翻訳機構と言語モデル機構のそれぞれの予測を元に翻訳を行う **Shallow Fusion** を提案した．この手法では，出力言語側の単言語コーパスで学習した言語モデル機構の予測確率を翻訳機構の予測確率と掛け合わせることで出力を決定する．学習においては事前に学習された言語モデルを用い，言語モデルのパラメータは固定した上で翻訳機構の学習を行う．そのため，言語モデルは翻訳の学習に影響されず，常にその言語らしい出力を予測する．

Shallow Fusion では，予測単語 $\hat{y}$ は

$$\hat{y} = \underset{y}{\operatorname{argmax}}(\log P_{\mathrm{TM}}(\mathbf{y}|\mathbf{x}) + \lambda \log P_{\mathrm{LM}}(\mathbf{y})) \tag{3.1.1}$$

によって表される．ここで，$\mathbf{x}$ は原言語文の入力，$P_{\mathrm{TM}}(\mathbf{y}|\mathbf{x})$ は翻訳機構の単語予測確率，$P_{\mathrm{LM}}(\mathbf{y})$ は言語モデル機構の単語予測確率を表す．また，λ は言語モデル機構を考慮する重みであり，人手で決められたパラメータである．なお，原論文では λ の値は $0.001 \leq \lambda \leq 0.1$ で設定しており，翻訳機構の情報を重視している．

Gulcehre らの実験では 5 つの設定（4 言語対）に対して実験を行い，ベースラインと比較して 1 つの設定はスコアが向上したが，その他の設定では同等もしくは悪化する結果となった．

3.2　Deep Fusion

Gulcehre ら [11] は，Shallow Fusion と同時に言語モデル機構に翻訳機構の情報を隠れ層で混ぜ合わせる **Deep Fusion** を提案した．このモデルでは言語モデル機構と翻訳機構は独立に事前学習しており、最終的に最後の予測部分だけを学習している．このモデルにおける言語モデルの考慮割合については言語モデルの隠れ層を元に決定されている．Shallow Fusion とは異なり，各機構は隠れ層同士を結合することで最終的な単語予測を行なっている．

Deep Fusion では，予測単語 $\hat{y}$ は

$$g = W_{\text{gate}} S_{\text{LM}}(\mathbf{y}) \tag{3.2.1}$$
$$h' = [S_{\text{TM}}(\mathbf{y}|\mathbf{x}); g \cdot S_{\text{LM}}] \tag{3.2.2}$$
$$S_{\text{deep}} = W_{\text{output}} h' \tag{3.2.3}$$
$$\hat{y} = \underset{y}{\text{argmax}}\, \text{softmax}(S_{\text{deep}}) \tag{3.2.4}$$

によって表される．ここで，$S_{\text{LM}}(\mathbf{y})$ と $S_{\text{TM}}(\mathbf{y}|\mathbf{x})$ は言語モデル機構と翻訳機構の隠れ層を表し，g は言語モデル機構の考慮割合を表す．また，W_{gate} $(|h| \times |h|)$，W_{output} $(2|h| \times |V|)$ はニューラルネットワークの重みを，そして $[a; b]$ は a と b の結合を表す．

Fulcehre らの実験では Shallow Fusion と同様の実験を行い，1 つの設定を除きスコアを向上させ最大で 2 ポイント程度の向上を見ることができた．

3.3 Cold Fusion

Sriram ら [15] は，Deep Fusion を発展させた手法として **Cold Fusion** を提案した．この手法では，翻訳機構と言語モデル機構の情報を考慮した gate 機構が用意されている．Deep Fusion との違いは大きく 2 点存在する．1 点目は言語モデル機構の使い方である．この手法では言語モデル機構の予測単語確率を求めたのちにもう一度隠れ層に戻して使用している．2 点目は gate 機構の考慮するものの違いである．この手法は言語モデル機構の情報を考慮する重みを，翻訳機構の隠れ層及び言語モデル機構の隠れ層の双方の情報を元に言語モデルの重みが決定される．これによって言語モデルの考慮する割合を翻訳の情報も用いながら決定している．

Cold Fusion では，予測単語 $\hat{y}$ は

$$h_{\text{LM}} = W_{\text{LM}} l_{\text{LM}} \tag{3.3.1}$$
$$g = W_{\text{gate}} [S_{\text{TM}}(\mathbf{y}|\mathbf{x}); h_{\text{LM}}] \tag{3.3.2}$$
$$h' = [S_{\text{TM}}(\mathbf{y}|\mathbf{x}); g \cdot h_{\text{LM}}] \tag{3.3.3}$$
$$S_{\text{cold}} = W_{\text{output}} h' \tag{3.3.4}$$
$$\hat{y} = \underset{y}{\text{argmax}}\, \text{softmax}(S_{\text{cold}}) \tag{3.3.5}$$

によって表される．ここで，l_{LM} は言語モデルのロジット化された単語予測確率

であり，h_{LM} は言語モデルの単語予測確率を改めて隠れ層に変換したものである．$S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x})$ は翻訳機構の隠れ層であり，g は言語モデル機構の考慮割合を表し，W_{LM} $(|h| \times |V|)$，W_{gate} $(2|h| \times |h|)$，W_{output} $(2|h| \times |V|)$ はニューラルネットワークの重みを表す．

この手法は音声認識の実験として投稿されているが，系列変換の手法であるため別タスクにも転用が可能である．次節で述べる Simple Fusion の比較手法として Stahlberg ら [12] が Cold Fusion の翻訳実験を報告している．その報告によると，4 言語対で実験を行い，Shallow Fusion と比較しても悪化する結果となった．なお，Sriram らが報告している音声認識の実験では 2% 程度（同一ドメインにおける単語誤り率での比較）の改善が見られている．

3.4 Simple Fusion

Stahlberg ら [12] は Cold Fusion を単純化した **Simple Fusion** を提案した．Cold Fusion と異なりこの手法では gate 機構を用いず固定した重みで混ぜ合わせる．また，言語モデル機構も複雑化せず同一のモデル内で完結する．

Simple Fusion には混ぜ合わせる方法の違う 2 つの似た手法である PostNorm（3.4.1）と PreNorm（3.4.2）を提案した．PostNorm と PreNorm における予測単語 $\hat{y}$ は，

$$\hat{y} = \underset{y}{\mathrm{argmax}}\, \mathrm{softmax}(\mathrm{softmax}(S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x})) \cdot P_{\mathrm{LM}}(\mathbf{y})) \tag{3.4.1}$$

$$\hat{y} = \underset{y}{\mathrm{argmax}}\, \mathrm{softmax}(S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x}) + \log P_{\mathrm{LM}}(\mathbf{y})) \tag{3.4.2}$$

によって表される．ここで，$S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x})$ は翻訳機構の隠れ層であり，$P_{\mathrm{LM}}(\mathbf{y})$ は言語モデル機構による単語予測確率である．

PostNorm では翻訳機構の単語予測確率と言語モデルの単語予測確率を掛け合わせることで最終的な単語を予測する．この時，それぞれの確率は重みを用いずに利用される．

PreNorm では言語モデル機構の単語予測対数確率と翻訳機構の正規化される前の単語予測確率[1]を足し合わせることで最終的な単語を予測する．なお，それぞ

[1] 正規化されていないため実際には確率ではないが便宜上確率と記述する。

れは重みを用いずに利用されているが，それぞれのスケールはそもそも違うことに留意されたい．

Simple Fusion モデルは比較的単純であるが，ベースラインや言語モデルを使用する他の手法と比較してもより高い BLEU を示すことが報告されており，Stahlberg らの行なった 4 言語対の実験において，言語モデルを用いないベースラインとの比較で約 0.4〜1.8 ポイントの向上が，言語モデルを用いる先行研究との比較で約 0.3〜1.6 ポイントの向上が示されている．なお，Stahlberg らによる設定において Shallow Fusion 及び Cold Fusion はベースラインとの比較でスコアの向上をほとんど見ることができず，また，Simple Fusion の 2 つの手法の優劣に一貫性は存在しない．

第 4 章　注意型言語モデルを用いたニューラル機械翻訳

4.1　提案手法：Dynamic Fusion

本論文では新たに Attention を用いた **Dynamic Fusion** を提案する．

Shallow Fusion や Simple Fusion においては事前に決められた重みに基づいて翻訳機構と言語モデル機構の情報を組み合わせている．しかしながら，翻訳において入力言語の情報を保持することは必要であり，単語単位で組み合わせる重みを調整すべきである．Cold Fusion では動的に重みを決定しているが、言語モデル機構の予測に対して処理を加えている．これは独立して学習された言語モデル機構に対して別の情報を与えてしまっているため，流暢性を重視した言語モデルが独立した予測をできなくなってしまっている．本手法では言語モデルは流暢性を高めるために補助的な情報を与えるため，翻訳機構とは独立に予測することとする．

さらに，先行研究においては単語予測確率を合わせることが多く，翻訳機構と言語モデル機構の語彙を統一する必要性がある．[1]一方で，本手法では attention として言語モデルを混ぜるため，翻訳機構と言語モデル機構の語彙が統一されていなくても全ての情報を扱うことができる．そのため，異なる単語区切り[2]や語彙圧縮手法[3]を用いることができ，一般に公開されている事前学習済み言語モデルでも用いることができる．

本手法では翻訳機構が単語 Attention を取り，その重みを言語モデル機構の単語予測確率と掛け合わせることで双方の情報を用いる．

まず，言語モデルの単語予測確率 $P_{\mathrm{LM}}(y)$ は

$$P_{\mathrm{LM}}(\mathbf{y}; y = \mathrm{word}) = \mathrm{softmax}(S_{\mathrm{LM}}(\mathbf{y})) \tag{4.1.1}$$

で表される．

[1]語彙が統一されない場合，どちらにも含まれる語彙以外が無視されてしまう
[2]語彙区切り：日本語のような単語境界が存在しない言語の文を単語に区切る手法
[3]語彙圧縮手法：扱う語彙数を減らすために行う手法．

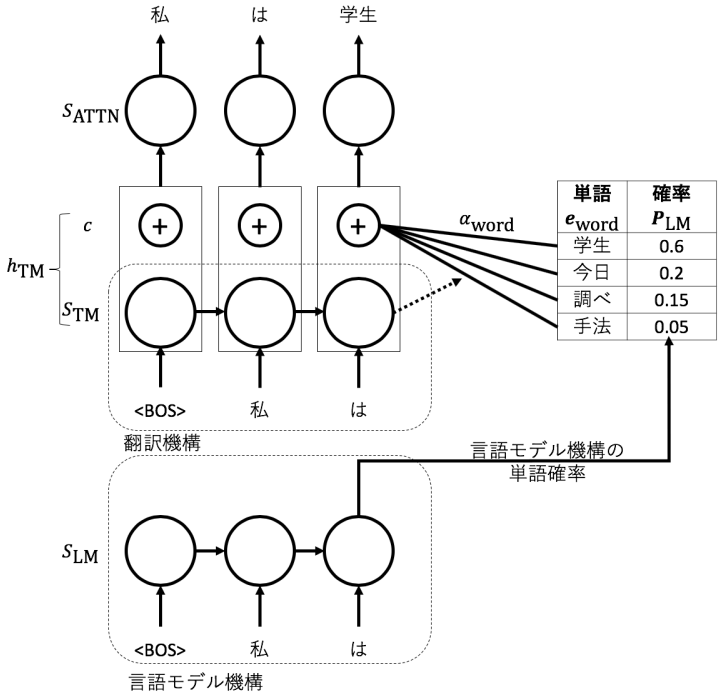

図 4.1　Dynamic Fusion の概略図.

次に，言語モデル機構に対して Attention を取った後の隠れ層 S_{ATTN} は

$$\alpha_{\mathrm{word}} = \frac{\exp(e_{\mathrm{word}}^{\mathrm{T}} S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x}))}{\sum_{\mathrm{word}\in V} \exp(e_{\mathrm{word}}^{\mathrm{T}} S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x})))} \tag{4.1.2}$$

$$c_{\mathrm{word}} = \alpha_{\mathrm{word}} e_{\mathrm{word}} \tag{4.1.3}$$

$$c_{\mathrm{LM}} = \sum_{\mathrm{word}} c_{\mathrm{word}} \cdot P_{\mathrm{LM}}(\mathbf{y}; y = \mathrm{word}) \tag{4.1.4}$$

$$h_{\mathrm{TM}} = [S_{\mathrm{TM}}(\mathbf{y}|\mathbf{x}); c_{\mathrm{LM}}] \tag{4.1.5}$$

$$S_{\mathrm{ATTN}} = W h_{\mathrm{TM}} \tag{4.1.6}$$

で表される．ここで e_{word} は単語埋め込みであり，c_{word} は各単語の従来の単語 Attention を表す．c_{LM} は単語 Attention の重みに言語モデル機構の予測確率を掛け合わせた後の Attention 隠れ層[4]を表し，W $(2|h| \times V)$ はニューラルネットワー

[4]単語 Attention は 2.2 節で述べた Luong らの Dot Attention を用いており，$\bar{h_i}$ の代わりに各単語埋め込み表現 e^{dec} を用いて計算される．なお，言語モデル機構の予測確率は α_{ij} に対して掛け合わされる．

表 4.1　先行研究と提案手法の比較.

手法	融合に用いるレイヤー	融合方法	融合割合（TM：LM）
Shallow Fusion	確率	積	人手で決定（固定）
Deep Fusion	隠れ層	結合	LM を元に gate 機構が自動で学習
Cold Fusion	確率 隠れ層相当に変換	結合	TM と LM を元に gate 機構が自動で学習
PostNorm	確率	積	重みなし（≒1：1）
PreNorm	隠れ層	和	重みなし（≒1：1）
Dynamic Fusion	LM：確率 TM：隠れ層	Attention	Attention が自動で学習

クの重みを表す.

式 (4.1.4) において c_{LM} は単語 Attention と $P_{\mathrm{LM}}(\mathbf{y}; y = \text{word})$ の内積である. 本手法において言語モデルの予測はそのトークン以前に出力された単語のみを考慮できる. 加えて，翻訳機構と言語モデル機構は従来の Attention 機構を用いて独立に計算された単語予測確率を合わせる.

最終的に予測単語 $\hat{y}$ は

$$\hat{y} = \underset{y}{\operatorname{argmax}} \operatorname{softmax}(S_{\mathrm{ATTN}}) \tag{4.1.7}$$

で表される.

Dynamic Fusion の概略図は図 4.1 に示す. このモデルでは、Attention を用いて言語モデル機構の情報を翻訳機構に使用する.

本手法の学習手順は Simple Fusion を踏襲して次の通りである：

1. 単言語コーパスで言語モデルを学習する.
2. 翻訳機構と Attention 機構を学習する.（言語モデル機構は固定する.）

4.2　先行研究と提案手法の比較

先行研究と提案手法には大きく 3 点で異なる点が存在する．その違いについて表 4.1 に示す．

1 点目は結合に用いるレイヤーである．先行研究では確率もしくは隠れ層を用いて同じレイヤーでの融合を行なっていたが，提案手法では異なるレイヤーを用いている．2 点目は結合の方法である．先行研究においては積・和・結合といった単純（重み付きを含む）に結合を行っていたのに対し，提案手法では Attention を用いた融合を行なっている．3 点目は融合割合の決定方法である．先行研究には固定されているものと自動で学習されるものがあり，提案手法では Attention 部分が自動で学習を行なっている．

第 5 章　実験

5.1　データ

5.1.1　コーパス

本論文では英語–日本語及び日本語–英語の 2 言語方向について実験を行う．本実験においては翻訳学習用のパラレルコーパスと言語モデル学習用の単言語コーパスの 2 つが必要である．そこで，the Asian Scientific Paper Excerpt Corpus (ASPEC) [18] を 2 つに分けて実験を行う．

この ASPEC は日本語で書かれた科学技術論文のアブストラクトとその翻訳である英語論文を文アライメントを用いて自動で関連づけたものである．そのため，文アライメントの確信度が低い文については翻訳になっていない文対も存在しており，学習データを全て使うことは一般に好まれない．そこで本研究では，学習データに含まれる約 300 万文対のうち自動スコアが高い 100 万文を翻訳学習用のパラレルコーパスとして利用し，残りの約 200 万文を言語モデル学習用の単言語コーパスとして用いる．なお，自動スコアが低い文対も各言語の文としては成立しており，単言語コーパスとしては用いることが可能である．

日本語文に対しては形態素解析器 MeCab[1] (IPADic) を用いて単語分割を行い，英語文に対しては Moses[2] (tokenizer, truecaser) を用いて前処理を行なった．また，処理の都合上学習データからは日本語文または英語文のどちらかが単語数が 60 を超える文対は削除し，開発データと評価データは公式のデータをそのまま利用した．実験に使用した文数は表 5.1 に示す．

5.1.2　語彙設定

語彙はパラレルコーパスのみを用いて作成され，頻度が高い順に 30,000 語を語彙として定義し，低頻度の語は OOV[3]として実験上未知語となる．そのため，単言

[1]https://github.com/taku910/mecab
[2]http://www.statmt.org/moses/
[3]OOV：ニューラルネットワークの計算量の都合上制限された語彙数に含めることのできない語彙

語コーパスのみに出現する単語は仮に頻出語であったとしても OOV となり，テスト時には未知語として扱われる．

加えて本実験では，語彙圧縮手法として広く用いられている Byte Pair Encoding (BPE) [19] を用いた実験も行う．BPE に関してはパラレルコーパスを基準に結合回数を 16,000 回に設定し，全ての語彙を使用する．なお，BPE は日本語と英語それぞれ別に適用する．

5.2 ベースライン及び比較手法

従来の AttentionNMT [4, 5] および Simple Fusion (PostNorm, PreNorm) を提案手法である Dynamic Fusion と比較する．ベースラインモデルとして Bahdanau ら [4] のモデルと Luong ら [5] のモデルを元に作成したモデル [20] を用い，他のモデルはこれを元に独自に作成したものを使用する．

5.3 パラメータ

比較のため全ての実験で同じパラメータを用いる．各設定については表 5.2 に示す．事前学習時には言語モデル機構のみを学習し，言語モデル機構を持たないベースラインではこの処理は実行しない．

5.4 評価方法

本論文では自動評価手法として，BLEU [16] および Rank-based Intuitive Bilingual Evaluation Score (RIBES) [17] を用いる．各実験につき 2 回ずつ実験を行い，その平均値をとる．また，ベースライン・Simple Fusion と Dynamic Fusion の間で Travatar [4]を用いて 10,000 回のブートストラップ法により有意差検定を行う．

[4]http://www.phontron.com/travatar/evaluation.html

表 5.1　コーパスの詳細.

	文（対）数	最大 トークン数
言語モデル用 (単言語コーパス)	1,909,981	60
学習用（対訳コーパス）	827,188	60
開発用（対訳コーパス）	1,790	
評価用（対訳コーパス）	1,812	

表 5.2　パラメータ設定.

	設定
事前学習 epoch 数	15 epoch
最大学習 epoch 数	100 epoch
最適化手法	AdaGrad
学習率	0.01
エンベディング次元数	512
隠れ層次元数	512
バッチ数	128
語彙数（w/o BPE）	30,000
BPE 結合回数	16,000

第 6 章　考察

6.1　自動評価に基づく分析

BLEU と RIBES による結果を表 6.1（英語–日本語）と表 6.2（日本語–英語）に示す．どちらのスコアも複数の設定を通じて一貫性が存在し，ベースラインと比較して Dynamic Fusion は BLEU と RIBES がおおよそ向上している．ベースラインと Simple Fusion を比較すると，ほとんどの設定で Simple Fusion はベースラインと同等もしくは悪化する結果となった．Simple Fusion と Dynamic Fusion を比較すると Dynamic Fusion は BLEU と RIBES がどちらも向上した．これらの向上から，提案手法が精度の向上に寄与しており，言語モデルを Attention で用いることが重要であることが示される．なお，英語–日本語の両言語対において Simple Fusion の手法を比較すると，PreNorm の精度が高いことがわかる．

英日翻訳において，言語モデルを用いることにより BLEU と RIBES が向上することが確認された．また，RIBES は Dynamic Fusion でさらに向上されており，提案手法が妥当性の高い出力をできていることがわかる．

ベースラインと Dynamic Fusion との間での BLEU と RIBES の有意差検定（$p < 0.05$）を行った.[1]その結果，日英翻訳の BPE を用いない設定では有意差が存在しなかったが，その他の設定においては有意差が確認された．これは，英語を目的言語にする翻訳は日本語が目的言語の場合と比較して，対訳コーパスのみでも文法知識を獲得できることによるものだと推察される．また，それに伴って先行研究では対訳コーパスから学習された文法知識に対して言語モデルがノイズとなってしまっているためにスコアが低下したものと考える．なお，Simple Fusion と Dynamic Fusion の間には日英翻訳では有意差が確認されたが，英日翻訳では確認されなかった．

加えてさらに現実的な設定として，翻訳モデルに BPE を適用し言語モデルには BPE を適用しない語彙設定で実験を行なった.[2]この設定においても BPE を用いたベースラインモデルと比較して翻訳精度の向上を得ることができた．

[1]有意差があったものについて表に*で示す．

[2]Simple Fusion においては語彙を揃える必要があるため実験を行うことはできない．

表 6.1　英日翻訳実験結果.

語彙設定	TM	w/o BPE		w/ BPE		w/ BPE	
	LM	w/o BPE		w/ BPE		w/o BPE	
		BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
ベースライン		31.28	80.78	32.35	81.17	32.35	81.17
PostNorm		31.01	80.77	32.43	80.97	N/A	N/A
PreNorm		31.61	80.78	32.69	81.24	N/A	N/A
Dynamic Fusion		**31.84***	**81.13***	**33.22***	**81.54***	**33.05***	**81.40***

表 6.2　日英翻訳実験結果.

語彙設定	TM	w/o BPE		w/ BPE		w/ BPE	
	LM	w/o BPE		w/ BPE		w/o BPE	
		BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
ベースライン		22.64	73.57	22.80	73.54	22.80	73.54
PostNorm		21.49	73.13	22.09	72.77	N/A	N/A
PreNorm		22.38	73.65	22.71	73.36	N/A	N/A
Dynamic Fusion		**22.78**	**73.74**	**23.45***	**74.01***	**23.08***	**73.73***

6.2　出力文を元にした定性的評価

それぞれのモデルの出力例を表 6.3 と 6.4 に示す.

表 6.3 の例において，ベースラインと比較して PreNorm と Dynamic Fusion の流暢性が向上した．加えて Dynamic Fusion では入力文中に存在する無生物主語を流暢に翻訳することができることがわかる．英語とは異なり，日本語において無生物主語はあまり使用されないため，英日翻訳において文字通りの翻訳は出力言語の母語話者にとって不自然に感じてしまう．なお，PostNorm は「線量」を「用量」と翻訳しており，妥当性が低下している.

表 6.4 の例において，PreNorm は単純かつ流暢な出力を示している．しかしながら，Simple Fusion の 2 つのモデルはベースラインと比較して妥当性を欠く出力となっている．対照的に，Dynamic Fusion では入力文の内容を参照訳以上に正しく翻訳している．これらから Dynamic Fusion は妥当性を失わずに流暢な出力を行うことができることが示される.

表 6.3　言語モデルによる流暢性向上例.

モデル	（出力）文
入力文	responding to these changes DERS can compute new dose rate .
参照訳	DERS は これら の 変化 に 対応 し て 新た な 線量 率 を 計算 できる 。
ベースライン	これら の 変化 に 対応 する 応答 は , 新しい 線量 率 を 計算 できる 。
Simple Fusion（PostNorm）	これら の 変化 に 対応 する 応答 は 新しい 用量 率 を 計算 できる 。
Simple Fusion（PreNorm）	これら の 変化 に 対応 する と , 新しい 線量 率 を 計算 できる 。
Dynamic Fusion	これら の 変化 に 対応 する こと により , 新しい 線量 率 を 計算 できる 。

表 6.4　Simple Fusion における妥当性の低下例.

モデル	（出力）文
入力文	the magnetic field is given in the direction of a right angle or a parallel (reverse to the flow) to the tube axis .
参照訳	磁場 は 管 軸 に 直角 か 平行 逆 方向 に 加え た 。
ベースライン	磁場 は 右 角 または 平行 (流れ) の 方向 に 与え られ , 管 軸 に 平行 で ある 。
Simple Fusion（PostNorm）	磁場 は 右 角度 または 平行 (流れ に 逆 に 逆) 方向 に 与え られ た 。
Simple Fusion（PreNorm）	磁場 は 右 角 または 平行 (流れ に 逆 方向) の 方向 に 与え られ た 。
Dynamic Fusion	磁場 は , 管 軸 に 直角 または 平行 (流れ に 逆 方向) の 方向 に 与え られる 。

これらの例から言語モデルの利用が出力の流暢性向上に寄与することを示し，さらに Dynamic Fusion では先行研究と比較して優れた妥当性を維持することが確認される．

日英翻訳では提案手法にとどまらず言語モデルを用いた全ての手法で表 6.5 のような態変化に対応できている．日本語で能動態を使用し英語側で受動態を使用する書き方は日本語の論文において一般的な方法であり [21]，示した例のように言語モデルを用いることでこの変化に対応できる．

6.3　言語モデルによる影響

言語モデルを用いることにより流暢性だけでなく妥当性を補う場合も存在する．表 6.6 はスペルミスの存在する入力文における例である．一般に入力文にスペル

表 6.5　態変化に対する頑健性向上例.

モデル	（出力）文
入力文	変形 が 対 密度 分布 に 影響 し て いる こと が 分かっ た 。
参照訳	it was found that the deformation gave effects to the pairing density distribution .
ベースライン	it was found that deformation was affected by the pair density distribution .
Simple Fusion (PostNorm)	it was found that deformation affects the logarithmic density distribution .
Simple Fusion (PreNorm)	it was found that deformation affected the pair density distribution .
Dynamic Fusion	it was found that the deformation affected the pair density distribution .

ミスが存在する場合，未知の単語であるため適切な翻訳ができない場合が存在する．この例では “temperature” を “temperture” と誤っているため，ベースラインモデルでは未知語として翻訳がなされてしまう．その結果として正しい翻訳が出力できなくなっている．しかし PreNorm と Dynamic Fusion では対応部分を補完し，妥当性を欠くことなく正しく翻訳することができる．これは言語モデルにより前後関係から補完すべき語を予測することができたことによるものであると考えられる．

6.4　Dynamic Fusion による影響

6.4.1　流暢性

Dynamic Fusion の出力と単語 Attention の重み（上位 5 単語の抜粋）を表 6.7 に示す.

文頭の語を除いて[3]，単語 Attention は確からしい出力を含んでいる．例えば鉤括弧（「）が文中に存在した場合，鉤括弧（」）で閉じようとする傾向が見られる．さらに，「発電」で鉤括弧を閉じることは望ましくないため，続く単語では「所」であると予測しており，それまでの出力を考慮した予測ができている．これは，入力文の情報を維持しながらも Dynamic Fusion が流暢性を改善することができること

[3]言語モデルは前の語までの情報から単語を予測するものであり，文頭記号（<BOS>）以外の情報がない文頭の語は予測が不可能である.

を示している．

単語 Attention の重みについては，複数の単語の重みが均一である場合と特定の単語に重みが大きく偏っている場合が存在する．主に機能語の出力など多くの翻訳方法が存在する場合に均一な重みを示すことが多い．この部分に関してはさらなる分析を行う必要があると考える．

6.4.2 妥当性

対照的に言語モデル機構が流暢性を犠牲にして妥当な翻訳を予測することは非常に稀である．そのため特定の単語の重みが他と比較してはるかに高い場合であっても，入力文の妥当性を損なう場合など Dynamic Fusion としての出力には使用されない場合がある．

実際，表 6.7 では文頭をはじめとして言語モデルの出力が考慮されていない単語出力が多く存在する．

この理由の 1 つに翻訳機構と言語モデル機構の貢献度の違いがあると考えられる．式（4.1.4）の変換重み行列を翻訳機構と言語モデル機構の行列に分解し，各行列の Frobenius ノルム[4]を計算した結果，翻訳機構は言語モデル機構の約 2 倍の貢献度であることがわかった．この点は Simple Fusion などの固定した重みで組み合わせる方法では扱うことができないものである．

6.4.3 言語モデルの役割

現在，既存の言語モデルのほとんどは入力文の情報を利用していない．したがって，言語モデルの流暢な予測によるノイズを排除するためには，言語モデル機構と翻訳機構は独立して予測を行い，翻訳機構からの Attention を考慮して用いられなければならない．ただし，言語モデルは出力言語における流暢な出力に関する情報を持っているという点で有益である．そのため，入力文がわからなくても流暢性の高い出力を出すことは可能である．

[4]Frobenius ノルム：行列の大きさを表す尺度の 1 つで，$\sqrt{\sum_{i,j} a_{i,j}^2}$ で計算される．

最終的に提案手法における言語モデル機構の役割は，翻訳機構が出力文の流暢性を改善するために出力言語の情報を補強することにある．そのため流暢性を向上させ妥当性を欠かない場合にのみ，言語モデル機構から翻訳のオプションを取得する．これは表 6.3 の例のように文体の微妙さを明確にする正則化方法とみなすこともできる．

表 6.6　言語モデルによる妥当性の比較.

モデル	（出力）文
入力文	this paper explains the application of chemical processes utilizing supercritical phase where a liquid does not make phase change irrespective of temperture or pressure .
参照訳	流体 が 温度 ・ 圧力 に かかわら ず 相 変化 し ない 状態 で ある 超 臨界 相 を 利用 し た 化学 プロセス の 応用 について 解説 し た 。
ベースライン	液体 が 相 変化 を 持た ない 超 臨界 相 を 利用 し た 化学 プロセス の 応用 について 解説 し た 。
Simple Fusion（PostNorm）	液体 が 相 変化 を 起こす こと なく , 圧力 や 圧力 に 関係 なく 相 変化 を 生じる 化学 プロセス の 適用 について 解説 し た 。
Simple Fusion（PreNorm）	液体 が 相 変化 を 起こさ ない 超 臨界 相 を 利用 し た 化学 プロセス の 応用 について , 温度 や 圧力 に 関係 なく 解説 し た 。
Dynamic Fusion	液体 が 温度 や 圧力 に 関係 なく 相 変化 を 起こさ ない 超 臨界 相 を 利用 し た 化学 プロセス の 応用 について 解説 し た 。

表 6.7　Dynamic Fusion による出力と単語 Attention の重み（抜粋）.

モデル	出力													
入力文	details of dose rate of " Fugen Power Plant " can be calculated by using <unk> software .													
参照訳	<unk> ソフトウエア を 用い て 「 ふ げん 発電 所 」 の 線量 率 を 詳細 に 計算 できる 。													
Dynamic Fusion	「 ふ げん 発電 所 」 の 線量 率 の 詳細 を , <unk> ソフトウェア を 用い て 計算 できる 。													
Dynamic Fusion（抜粋）	「		ふ		げん		発電		所		」		の	
単語 Attention（上位 5 語）の重み	本	9.9e-1	この	5.5e-1	」	9.9e-1	」	1.0	所	9.9e-1	」	1.0	について	7.7e-1
	標記	8.7e-5	その	3.5e-1	ね	3.2e-6	号	2.7e-8	機	1.3e-4	発電	3.2e-12	の	1.7e-1
	この	4.2e-5	日本	7.0e-2	げん	2.0e-9	げん	1.4e-11	」	1.2e-6	の	1.7e-18	における	4.5e-2
	また	8.5e-6	1	2.7e-2	出	1.1e-10	<unk>	1.1e-12	設備	7.7e-11	<unk>	7.6e-19	で	6.4e-3
	これら	1.5e-6	高	4.7e-3	り	3.6e-11	・	1.8e-14	装置	2.6e-12	用	6.3e-19	と	3.2e-3

第 7 章　おわりに

本論文ではニューラル機械翻訳に対して言語モデルを活用する新たなモデルである Dynamic Fusion を提案した．翻訳実験によって英日翻訳において言語モデルを用いることはスコアの向上に寄与し，提案手法においては日英翻訳においてもスコアが向上することを示した．また先行研究との比較によって，言語モデルを単純に融合させるのではなく提案手法のように注意機構的に用いることが有用であることを示した．言語モデル機構と翻訳機構の重みを固定する簡易性と比較しても，注意機構は言語モデルを用いて妥当性を損なうことなく流暢性を向上させることに寄与していると言える．これは BLEU や RIBES のさらなる向上が得られ，質の高い翻訳を得ることができる．

加えて，提案手法においては翻訳機構と言語モデル機構の語彙を一致させる必要性が必ずしもあるわけではなく，本研究においては BPE の有無で語彙を区別した設定でもスコアが向上することが確認された．これにより一般に公開されている学習済みの言語モデルを用いることなども可能となることが推察され，より広範的に活用できることが期待される．

本手法は注意機構によって各単語の重みは自動で調整されるが，注意機構自身の翻訳に与える重みに関しては固定されている．将来的には文脈などに応じて言語モデルと翻訳モデルの融合する重み（寄与度）に関しても各出力ステップごとに自動調整が行える機構について検討されたい．

謝辞

本論文の執筆に際して，研究室に配属されてからの 3 年間にわたり研究に関してご指導いただきました小町守准教授に深く感謝いたします．また，国際会議をはじめとした多くの発表機会，メンターとしての後輩の指導や企業の方との共同研究など，様々な経験をさせていただきまして感謝しております．

研究室配属当初からお世話になりました松村雪桜さんと山岸駿秀さんには，研究の進め方から論文の執筆までアドバイスを頂き感謝しております．

そして，山口亨教授，高間康史教授には副査を引き受けて頂き大変感謝しております．

最後に，研究生活において多岐にわたりお世話になりました研究室の皆さま，インターンシップや共同研究でお世話になりました企業の方々，学会などで様々なアドバイスをいただいた方々など，研究生活において関わった皆様に感謝いたします．

参考文献

[1] I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to Sequence Learning with Neural Networks,” Advances in Neural Information Processing Systems 27, eds. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, pp.3104–3112, Curran Associates, Inc., 2014. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[2] K. Cho, B. vanMerriënboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches,” Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp.103–111, Association for Computational Linguistics, Doha, Qatar, Oct. 2014. https://www.aclweb.org/anthology/W14-4012

[3] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, “Modeling Coverage for Neural Machine Translation,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.76–85, Association for Computational Linguistics, Berlin, Germany, Aug. 2016. https://www.aclweb.org/anthology/P16-1008

[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” Proceedings of the 3rd International Conference on Learning Representations, pp.1–15, San Diego, California, America, May 2015. https://arxiv.org/abs/1409.0473

[5] T. Luong, H. Pham, and C.D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.1412–1421, Association for Computational Linguistics, Lisbon, Portugal, Sept. 2015. https://www.aclweb.org/anthology/D15-1166

[6] T. Brants, A.C. Popat, P. Xu, F.J. Och, and J. Dean, “Large Language Models in Machine Translation,” Proceedings of the 2007 Joint Confer-

ence on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp.858–867, Association for Computational Linguistics, Prague, Czech Republic, Jun. 2007. https://www.aclweb.org/anthology/D07-1090

[7] P. Ramachandran, P. Liu, and Q. Le, "Unsupervised Pretraining for Sequence to Sequence Learning," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.383–391, Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017. https://www.aclweb.org/anthology/D17-1039

[8] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?," Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp.529–535, Association for Computational Linguistics, New Orleans, Louisiana, Jun. 2018. https://www.aclweb.org/anthology/N18-2084

[9] T. Hirasawa, H. Yamagishi, Y. Matsumura, and M. Komachi, "Multimodal Machine Translation with Embedding Prediction," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp.86–91, Association for Computational Linguistics, Minneapolis, Minnesota, Jun. 2019. https://www.aclweb.org/anthology/N19-3012

[10] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.86–96, Association for Computational Linguistics, Berlin, Germany, Aug. 2016. https://www.aclweb.org/anthology/P16-1009

[11] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On Using Monolingual Corpora in Neural Machine Translation," 2015. https://arxiv.org/abs/1503.03535

[12] F. Stahlberg, J. Cross, and V. Stoyanov, "Simple Fusion: Return of the Language Model," Proceedings of the Third Conference on Machine Translation: Research Papers, pp.204–211, Association for Computational Linguistics, Brussels, Belgium, Oct. 2018. https://www.aclweb.org/anthology/W18-6321

[13] R. Sennrich and B. Haddow, "Linguistic Input Features Improve Neural Machine Translation," Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, pp.83–91, Association for Computational Linguistics, Berlin, Germany, Aug. 2016. https://www.aclweb.org/anthology/W16-2209

[14] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.2126–2136, Association for Computational Linguistics, Melbourne, Australia, Jul. 2018. https://www.aclweb.org/anthology/P18-1198

[15] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold Fusion: Training Seq2Seq Models Together with Language Models," 2017. https://arxiv.org/abs/1708.06426

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.311–318, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, Jul. 2002. https://www.aclweb.org/anthology/P02-1040

[17] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic Evaluation of Translation Quality for Distant Language Pairs," Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp.944–952, Association for Computational Linguistics, Cambridge, MA, Oct. 2010. https://www.aclweb.org/anthology/D10-1092

[18] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kuro-

hashi, and H. Isahara, “ASPEC: Asian Scientific Paper Excerpt Corpus,” Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), eds. by N.C.C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, pp.2204–2208, European Language Resources Association (ELRA), Portoroz, Slovenia, May 2016.

[19] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1715–1725, Association for Computational Linguistics, Berlin, Germany, Aug. 2016. https://www.aclweb.org/anthology/P16-1162

[20] Y. Matsumura and M. Komachi, “Tokyo Metropolitan University Neural Machine Translation System for WAT 2017,” Proceedings of the 4th Workshop on Asian Translation (WAT2017), pp.160–166, Asian Federation of Natural Language Processing, Taipei, Taiwan, Nov. 2017. https://www.aclweb.org/anthology/W17-5716

[21] H. Yamagishi, S. Kanouchi, T. Sato, and M. Komachi, “Improving Japanese-to-English Neural Machine Translation by Voice Prediction,” Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp.277–282, Asian Federation of Natural Language Processing, Taipei, Taiwan, Nov. 2017. https://www.aclweb.org/anthology/I17-2047

発表リスト

国際会議

1. Michiki Kurosawa, Yukio Matsumura, Hayahide Yamagishi and Mamoru Komachi. **Japanese Predicate Conjugation for Neural Machine Translation.** In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (NAACL2018–SRW), pp.100–105. New Orleans, Louisiana, USA. Jun, 2018.
2. Michiki Kurosawa and Mamoru Komachi. **Dynamic Fusion: Attentional Language Model for Neural Machine Translation.** In Proceedings of the 2019 16th International Conference of the Pacific Association for Computational Linguistics (PACLING2019), #18, pp.1–13. Hanoi, Vietnam. October, 2019.

国内会議

1. 黒澤 道希, 山岸 駿秀, 松村 雪桜, 小町 守. **活用情報を用いた日英ニューラル機械翻訳.** NLP 若手の会第 12 回シンポジウム (YANS2017). 那覇. 2017 年 9 月.
2. 黒澤 道希, 松村 雪桜, 山岸 駿秀, 小町 守. **述語の活用情報を用いたニューラル日英翻訳.** 言語処理学会第 24 回年次大会 (NLP2018), pp.813–816. 岡山. 2018 年 3 月.
3. 黒澤 道希, 小町 守. **ニューラル機械翻訳に対する言語モデルの導入に関する検討.** NLP 若手の会第 13 回シンポジウム (YANS2018). 高松. 2018 年 8 月.
4. 長我部恭行, 甲斐優人, 石井奏人, 荻野天翔, 黒澤道希, 小町守. **機械翻訳に対する文間文脈を考慮した評価と分析.** 言語処理学会第 25 回年次大会 (NLP2019), pp.1073–1076. 名古屋. 2019 年 3 月.

5. 黒澤 道希, 小町 守. **ニューラル機械翻訳に対する注意言語モデル.** 情報処理学会第 240 回自然言語処理研究会 (IPSJ–NL; NL 研), Vol.2019–NL–240 No.13, pp.1–6. 遠野. 2019 年 1 月.
6. 中澤真人, 嶋中宏希, 黒澤道希, 小町守. **中日機械翻訳における事前学習された言語モデリングの利用に関する考察.** NLP 若手の会第 14 回シンポジウム (YANS2019). 札幌. 2019 年 8 月.