# 翻訳における言語モデルとしての High order Joint Probability の分野依存性の調査

松本大輝 [*1]　村上仁一 [*2]
[*1] 鳥取大学 工学部 電気情報系学科
[*2] 鳥取大学大学院　工学研究科　情報エレクトロニクス専攻
s152100@ike.tottori-u.ac.jp[*1]
murakami@eecs.tottori-u.ac.jp[*2]

## 1　はじめに

現在機械翻訳の多くの言語モデルには $N$-gram モデルが使用されている．しかし，$N$-gram は分野依存性があり，分野の違う言語データを追加しても翻訳精度があまり変わらないことが知られている [1].

そこで本研究では，$N$-gram モデルの代わりに交差エントロピーの拡張である High order Joint Probability を用いる．$N$-gram が近接の情報を利用しているのに対し，High order Joint Probability は遠隔の情報を利用できる．そのため分野依存性が少なく，分野の違う言語データを追加しても翻訳精度が向上すると考える．

## 2　$N$-gram（従来手法）

一般的な手法では，言語モデルに $N$-gram モデルを使用している．以下の式 (1) に本研究で使用する Tri-gram の計算式を示す．

$$\sum_{i=0}^{N-1} \log_2 \left( \frac{count(E_{i-2}, E_{i-1}, E_i)}{count(E_{i-2}, E_{i-1})} \right) \tag{1}$$

$E_i$：出力英語単語　$N$：英文の単語数　$count$：対訳学習文の頻度

## 3　従来手法の問題点

Trigram は近隣の情報を利用している．そして遠隔の情報を利用せず，名詞と動詞の関係などを利用しない．そのため分野依存性が高くなる．この問題を解決するために，本実験では異なる言語モデルとして High order joint Probability を提案する．

## 4　High order Joint Probability（提案手法）

本研究では，言語モデルに $N$-gram モデルの代わりに High order Joint Probability（以下 Joint Probability）を用いる．以下の式 (2) にその計算式を示す [2].

$$\sum_{j=0}^{M-1} \sum_{i=0}^{N-1} P(J_{j-2}, J_{j-1}, J_j, E_{i-2}, E_{i-1}, E_i)$$

$$\times \log_2 \frac{P(J_{j-2}, J_{j-1}, J_j, E_{i-2}, E_{i-1}, E_i)}{P(J_{j-2}, J_{j-1}, J_j) P(E_{i-2}, E_{i-1}, E_i)}$$

$$= \sum_{j=0}^{M-1} \sum_{i=0}^{N-1} \frac{count(J_{j-2}, J_{j-1}, J_j, E_{i-2}, E_{i-1}, E_i)}{N_{je}}$$

$$\times \log_2 \frac{\frac{count(J_{j-2}, J_{j-1}, J_j, E_{i-2}, E_{i-1}, E_i)}{N_{je}}}{\frac{count(J_{j-2}, J_{j-1}, J_j)}{N_j} \frac{count(E_{i-2}, E_{i-1}, E_i)}{N_e}} \tag{2}$$

$J_j$：日本語単語　$M$：日本文の単語数
$E_i$：英語単語　$N$：英文の単語数
$P$：出現確率　$count$：対訳学習文中の頻度
$N_{je}$：対訳学習文中の日本語単語の三つ組と英語単語の三つ組の総数
$N_j$：対訳学習文中の日本語単語の三つ組の総数
$N_e$：対訳学習文中の英語単語の三つ組の総数

## 5　提案手法の利点

Trigram は近隣の単語の情報を使用している．しかし遠隔の情報が翻訳に必要な場合がある．入力文を表 1 に，出力候補文を表 2 に示す．近隣の単語の情報を使用する Tri-gram は，候補文 1 では「candle」が「flickering」の関係性を，候補文 2 では「candle」が「swinging」の関係性を用いる．しかし，この近接の単語情報だけでは，表 2 中の出力候補文の選択が困難である．

一方，High order Joint Probability は遠隔の日英単語の情報を使用できる．このため入力文中の日本語単語「火」に対して候補文 1 では「flickering」，候補文 2 では「swinging」との関係性を使用することができる．この遠隔の情報を使用することで適切な候補文 1 を選択することができると考える．

表 1　入力文

| | |
|---|---|
| 入力文 | ろうそくの火が揺れている。 |

表 2　出力候補文

| | |
|---|---|
| 候補文 1 | The flame of the candle is flickering . |
| 候補文 2 | The flame of the candle is swinging . |

## 6　実験

### 6.1　実験データ

表 3 に実験データを示す．対訳学習文には辞書文で単文 [3] を使用する．テスト文も辞書文で単文を使用する．追加データには、学習文とは別分野のデータである wikipedia 文 [4] を使用する．表 4 に対訳学習文と wikipedia 文の平均単語長を示す．表 5 にテスト文と wikipedia 文の例を示す．

表 3　実験データ

| | |
|---|---|
| 対訳学習文（辞書文） | 159,998 文 |
| テスト文（辞書文） | 100 文 |
| 追加データ（wikipedia 文） | 343,308 文 |

表 4　平均単語長

| | |
|---|---|
| 対訳学習文 | 8.7 |
| wikipedia 文 | 29.4 |

表 5　データ例

| | |
|---|---|
| テスト文（辞書文） | |
| 日文 1 | ピアノ の 勉強 に ヨーロッパ に 行く 。 |
| 英文 1 | Go to Europe to study the piano . |
| 日文 2 | 1 平方 メートル 20 万 円 で 彼 と その 商談 を 取り決め た 。 |
| 英文 2 | I stuck the bargain with him at two hundred thousand yen a square meter . |
| wikipedia 文 | |
| 日文 1 | 彼 が 大統領 に 選ば れる 公算 が 大きい 。 |
| 英文 1 | There is every probability that he will be elected President . |
| 日文 2 | 彼 ら に とって は 幹部 に 認めて もらう 機会 で あり 、 腕 の 見せ所 と なって いる 。 |
| 英文 2 | Playing a ” mitate ” is an opportunity for them to show their skills and get recognition from high-ranking people . |

## 6.2　評価方法

本実験では翻訳モデルに相対的意味論に基づく変換主導型パターンベース統計機械翻訳（以下 TDPBSMT）[5] を用いる．言語モデルは $N$-gram と Joint Probability を利用する．対訳学習文には辞書文で単文を用いる．テスト文も辞書文で単文を用いる．追加データには wikipedia 文を用いる．翻訳の評価として自動評価と人手評価を用いる．

# 7　実験結果

## 7.1　自動評価結果

表 6 に自動評価結果を示す．

表 6　自動評価結果（wikipedia 文）

| wikipedia 文追加 | BLEU | METEOR | RIBES | TER |
|---|---|---|---|---|
| Joint Probability | **0.09** | **0.38** | **0.72** | **0.72** |
| Tri-gram | 0.03 | 0.29 | 0.68 | 0.74 |

## 7.2　人手評価結果

得られた出力結果を人手にて評価した．評価基準を以下に示す．また，人手評価結果を以下の表 7 に示す．

- Joint Probability ○ ： Joint Probability の結果が Tri-gram よりも良い
- Tri-gram ○ ： Tri-gram の結果が Joint Probability よりも良い
- 差なし ： Tri-gram と Joint Probability の翻訳精度に差がない
- 同一 ： Tri-gram と Joint Probability の出力結果が同じ
- 未出力 ： Tri-gram と Joint Probability の出力結果が共に未出力

表 7　人手評価結果

| Joint Probability ○ | Tri-gram ○ | 差なし | 同一 | 未出力 |
|---|---|---|---|---|
| 23 | 2 | 58 | 12 | 5 |

表 7 における Joint Probability ○の出力例を表 8 と 9 に示す．

表 8　Joint Probability ○の出力例 1

| | |
|---|---|
| 入力文 | 中国人 が 印刷 技術 を 発明 し た 。 |
| 参照文 | The Chinese invented printing . |
| Tri-gram | The **printing** invented the Chinese . |
| Joint Probability | The Chinese invented the printing . |

表 9　Joint Probability ○の出力例 2

| | |
|---|---|
| 入力文 | 公園 は 川 まで 広がっ て いる 。 |
| 参照文 | The park reaches to the river . |
| Tri-gram | **park** spreading river . |
| Joint Probability | The park is spread to the river . |

表 7 における Tri-gram ○の出力例を表 10 と表 11 に示す．

表 10　Tri-gram ○の出力例 1

| | |
|---|---|
| 入力文 | 鞄 が 肩 に 掛かっ て いる 。 |
| 参照文 | The bag is hung over his shoulders . |
| Tri-gram | The bag hanging shoulder . |
| Joint Probability | He is the 掛かっ bag shoulders . |

表 11　Tri-gram ○の出力例 2

| | |
|---|---|
| 入力文 | 時 が 彼女 の 悲しみ を いやし た 。 |
| 参照文 | Time healed her sorrow . |
| Tri-gram | Time healed She grief . |
| Joint Probability | She Time healed **his** grief . |

## 7.3　評価結果のまとめ

表 6 と表 7 の結果より，対訳学習文とは別分野のデータを追加データとして用いた場合，Joint Probability の方が Tri-gram よりも良いことが分かる．特に人手評価については大きな差が見られた．

# 8　考察

## 8.1　追加前

Tri-gram と Joint Probability の性能を比較するため，表 3 において追加データを利用せずに実験を行った．表 12 に実験データを，表 13 に自動評価結果を，表 14 に人手評価結果を示す．

表 12　実験データ

| | |
|---|---|
| 対訳学習文 | 159,998 文 |
| テスト文 | 100 文 |

表 13　自動評価結果（追加前）

| | BLEU | METEOR | RIBES | TER |
|---|---|---|---|---|
| Joint Probability（追加前） | **0.12** | **0.41** | **0.72** | **0.69** |
| Tri-gram（追加前） | 0.10 | 0.38 | 0.71 | 0.71 |

表 14　人手評価結果（追加前）

| Joint Probability ○ | Tri-gram ○ | 差なし | 同一 | 未出力 |
|---|---|---|---|---|
| 14 | 6 | 48 | 27 | 5 |

表 14 における Joint Probability ○の出力例を表 15 と表 16 に示す．

表 15　Joint Probability ○の出力例 1

| | |
|---|---|
| 入力文 | この 言葉 は 一般 に 通用 し て いる 。 |
| 参照文 | The words are in common currency . |
| Tri-gram | This word is **not** accepted for the public . |
| Joint Probability | This word is not accepted for the public . |

表 16　Joint Probability ○の出力例 2

| | |
|---|---|
| 入力文 | 彼女 の 結婚 は 一大 行事 と なる だろ う 。 |
| 参照文 | Her marriage will be a great event . |
| Tri-gram | **She** will be a great event in marriage . |
| Joint Probability | Her marriage will be a great event . |

表 14 における Tri-gram ○の出力例を表 17 と表 18 に示す．

表 17　Trigram ○の出力例 1

| | |
|---|---|
| 入力文 | 鞄 が 肩 に 掛かっ て いる 。 |
| 参照文 | The bag is hung over his shoulders . |
| Tri-gram | the bag is hanging on the shoulders . |
| Joint Probability | He is 掛かっ shoulders of the bag . |

表 18　Trigram ○の出力例 2

| | |
|---|---|
| 入力文 | 幕 が だらり と 垂れ下がっ て いる 。 |
| 参照文 | The curtain hangs down loosely . |
| Tri-gram | lolled curtain is hanging down . |
| Joint Probability | The curtain is drooping **tongue** . |

表 13 と表 14 の結果より，追加前においても，Joint-Probability の方が Tri-gram よりも良いことが分かる．

## 8.2　同分野（複文）

対訳学習文と同分野のデータを追加した場合の比較を行った．表 19 に実験データを示す．言語モデルには Tri-gram と Joint Probability を使用した．追加データには辞書文から抽出した複文を利用した．表 20 に学習文と複文の平均単語長を示す．表 21 に複文の例を示す．表 22 に自動評価結果を示す．表 23 に人手評価結果を示す．

表 19　実験データ

| | |
|---|---|
| 対訳学習文（辞書文の単文） | 159,998 文 |
| テスト文（辞書文の単文） | 100 文 |
| 追加データ（辞書文の複文） | 100,000 文 |

表 20　平均単語長

| | |
|---|---|
| 対訳学習文 | 8.7 |
| 複文 | 11.3 |

表 21　データ例

| | 複文 |
|---|---|
| 日文 1 | そんな 事 が あっ た と は 少し も 知り ませ ん でし た 。 |
| 英文 1 | I did not know at all that there was such a thing . |
| 日文 2 | 貴社 から の これまで の ご 援助 に お礼 を 申し上げ ます とともに 、両社 の 素晴らしい 関係 が 今後 とも 続き ます こと を 期待 し て おり ます 。 |
| 英文 2 | I also want to thank you for your past support and I am looking forward to continued excellent relations between our companies . |

表 22　自動評価結果（複文）

| 複文追加 | BLEU | METEOR | RIBES | TER |
|---|---|---|---|---|
| Joint Probability | **0.15** | **0.42** | **0.75** | **0.68** |
| Tri-gram | 0.11 | 0.38 | 0.69 | 0.71 |

表 22 の結果より，同一分野のデータを追加した場合，自動評価結果は Joint Probability の方が Tri-gram よりも良いことが分かる．

表 23　人手評価結果（複文）

| Joint Probability ○ | Tri-gram ○ | 差なし | 同一 | 未出力 |
|---|---|---|---|---|
| 17 | 6 | 45 | 27 | 5 |

表 23 における Joint Probability ○の出力例を表 24 と表 25 に示す．

表 24　Joint Probability ○の出力例 1

| | |
|---|---|
| 入力文 | 溶岩 が 少し 後退 し た 。 |
| 参照文 | The lava has retreated a little . |
| Tri-gram | **little** lava retreated . |
| Joint Probability | The lava retreated little . |

表 25　Joint Probability ○の出力例 2

| | |
|---|---|
| 入力文 | 私 は その 仕事 に 慣れ て い ない 。 |
| 参照文 | I am not used to the task . |
| Tri-gram | I **was** not used to the job . |
| Joint Probability | I am not used to the work . |

表 23 における Tri-gram ○の出力例を表 26 と表 27 に示す．

表 26　Tri-gram ○の出力例 1

| | |
|---|---|
| 入力文 | この 言葉 は 一般 に 通用 し て いる 。 |
| 参照文 | The words are in common currency . |
| Tri-gram | This word is spoken in public . |
| Joint Probability | This word has **accepted** the public . |

表 27　Tri-gram ○の出力例 2

| | |
|---|---|
| 入力文 | 彼 の 性質 は ひねくれ て いる 。 |
| 参照文 | He has an uneven disposition . |
| Tri-gram | His character is twisted . |
| Joint Probability | He **has** distorted in the nature of his . |

表 22 と表 23 の結果より，対訳学習文と同分野のデータを追加データとした場合，Joint Probability の方が Tri-gram よりも良いことが分かる．

## 8.3　評価結果のまとめ

どの場合においても Joint Probability の方が Tri-gram よりも良かった．同一分野の言語データを追加した場合でも，Joint Probability の方が Tri-gram よりも良かった．他分野の言語データを追加した場合では，特に Joint Probability が有効であることが分かった．

## 8.4 解析

TDPBSMT は確率を用いて最終的な出力文を決定する．その計算式を以下に示す．

$$\log_2 P = \log_2 P_v + \log_2 P_p + \log_2 P_m \quad (3)$$

$P$：翻訳確率
$P_v$：対訳学習文中の単語の出現回数に基づく確率
$P_p$：文パターンに付与された確率
$P_m$：言語モデルで生成された確率

### 8.4.1 Joint Probability ○の解析

表 8 における Tri-gram の出力候補文の翻訳確率を表 28 に，Joint Probability の出力候補文の翻訳確率を表 29 に示す．

表 28　出力候補文の翻訳確率（Tri-gram）

| 出力候補文 | The printing invented the Chinese . |
|---|---|
| $\log_2 P_v$ | -16.4 |
| $\log_2 P_p$ | -7.64 |
| $\log_2 P_m$（Tri-gram） | -71.8 |
| $\log_2 P_m$（Joint Probability） | -2539.3 |
| $\log_2 P$ | -95.8 |

表 29　出力候補文の翻訳確率（Joint Probability）

| 出力候補文 | The Chinese invented the printing . |
|---|---|
| $\log_2 P_v$ | -13.2 |
| $\log_2 P_p$ | -5.94 |
| $\log_2 P_m$（Tri-gram） | -90.9 |
| $\log_2 P_m$（Joint Probability） | -2474.7 |
| $\log_2 P$ | -2493.8 |

表 28 と表 29 より，$\log_2 P_m$（Tri-gram）の値は表 28 の方が表 29 よりも大きいことが分かる．また，$\log_2 P_m$（Joint Probability）の値は，表 29 の方が表 28 よりも大きいことが分かる．

### 8.4.2 Tri-gram ○の解析

表 10 における Tri-gram の出力候補文の翻訳確率を表 30 に，Joint Probability の出力候補文の翻訳確率を表 31 に示す．

表 30　出力候補文の翻訳確率（Tri-gram）

| 出力候補文 | The bag is hung over his shoulders . |
|---|---|
| $\log_2 P_v$ | -18.6 |
| $\log_2 P_p$ | -4.80 |
| $\log_2 P_m$（Tri-gram） | -82.8 |
| $\log_2 P_m$（Joint Probability） | -2082.7 |
| $\log_2 P$ | -106.2 |

表 31　出力候補文の翻訳確率（Joint Probability）

| 出力候補文 | He is the 掛かっ bag shoulders . |
|---|---|
| $\log_2 P_v$ | -147.6 |
| $\log_2 P_p$ | -6.86 |
| $\log_2 P_m$（Tri-gram） | -91.7 |
| $\log_2 P_m$（Joint Probability） | -1813.6 |
| $\log_2 P$ | -1968.1 |

表 30 と表 31 より，$\log_2 P_m$（Tri-gram）の値は表 30 の方が表 28 よりも大きいことが分かる．また，$log_2 P_m$（Joint Probability）の値は，表 31 の方が表 30 よりも大きいことが分かる．

### 8.4.3 解析の評価

8.4.1 項と 8.4.2 項の結果の比較より，Joint probability ○の翻訳確率と Tri-gram ○の翻訳確率に同様の傾向が見られた．Tri-gram ○と判断したもう 1 文についても同様の傾向が見られた．この事は方式の限界を示している．また，$P_v$ を見ると言語モデル無しで正しい出力候補文を選択できる可能性がある．今後は他にも様々な言語モデルがあるので，それらを用いて実験を行いたいと考えている．

# 9 おわりに

本研究では，言語モデルに一般的に使用される Tri-gram の代わりに Joint Probability を使用する手法を提案した．実験結果より，学習データと分野の異なるデータを追加した場合，Joint Probability では翻訳精度が大きく向上することが確認できた．今後は他にも様々な言語モデルがあるので，それらを用いて実験を行いたいと考えている．

# 参考文献


[1] 善行　佑介，村上 仁一，徳久 雅人，“モノリンガルデータを増加させた場合の統計的機械翻訳の精度調査”，2014.
[2] X.D.HAUNG，“HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION，p.49
[3] 村上　仁一，藤波　進，“日本語と英語の対訳文対の収集と著作権の考察”，第一回コーパス日本語学会ワークショップ，pp.119-130，2012.
[4] Wikipedia 日英京都関連文書対訳コーパス，https://alaginrc.nict.go.jp/WikiCorpus/
[5] 中村　勇太，村上　仁一，“パターンに基づく変換主導型統計機械翻訳（TDPBSMT）の提案”，2018.