

統計的機械翻訳におけるフレーズ対応最適化を利用した N-best 翻訳候補のリランキング

越川 満^{†1} 内山 将夫^{†2} 梅谷 俊治^{†3}
松井 知己^{†4} 山本 幹雄^{†1}

フレーズベース統計的機械翻訳では、連続する単語列（フレーズ）を翻訳の最小単位とした確率的規則に基づいて翻訳候補の順位付けを行い、最も確率の高い候補を出力とする。しかし、入力文のフレーズ区切りや翻訳前後の訳語関係（フレーズ対応）の組合せ数は膨大である。そのため、従来の統計的機械翻訳システムは、翻訳候補およびフレーズ区切り・対応に対して大胆な近似を行うことで探索空間を狭めており、厳密な確率の最大化をしていない。本稿では、フレーズ対応・区切りに関する厳密な確率最大化を行う問題を、フレーズベース翻訳において広く用いられているすべての素性を考慮可能な形式で整数線形計画問題として定式化し、それを翻訳候補のリランキングに応用する手法を提案・実装する。評価実験の結果、提案手法は有意に翻訳精度を改善することが示されると同時に、フレーズベース翻訳における探索の課題は、フレーズ対応ではなく翻訳候補文についてより多くの候補を評価することにあるという示唆が得られた。

N-best Reranking Using Optimal Phrase Alignment for Statistical Machine Translation

MITSURU KOSHIKAWA,^{†1} MASAO UTIYAMA,^{†2}
SHUNJI UMETANI,^{†3} TOMOMI MATSUI^{†4}
and MIKIO YAMAMOTO^{†1}

Phrase-based statistical machine translation system outputs the candidate having the highest probability based on the probabilistic phrase translation rules. However, there exist a huge number of translation candidates and ambiguities on phrase segmentations/alignments for source and target sentences. Therefore, the current statistical translation systems use various heuristics for reducing the number of translation candidates and approximating phrase-alignment probabilities, in order to narrow the search space. This paper pro-

poses the formulation to strictly maximize the phrase-alignment probability computed from all features which most phrase-based statistical machine translation systems use within. We also propose a reranking method based on the proposed phrase alignment optimization. In evaluation experiments, our system improved significantly the translation quality. The experimental results also suggested that a variety of translation candidates are more important for increasing accuracy than exact phrase alignments.

1. はじめに

インターネットが普及して以来、ウェブ上に存在する情報は増加の一途をたどっている。この情報を最大限に利用するには、様々な言語で書かれたテキスト情報を翻訳する必要があるが、人手によりすべてのウェブページを翻訳することは非現実的である。そこで、翻訳の手段の1つとして機械翻訳が期待されている。

統計的機械翻訳では、対訳コーパスと呼ばれる同一の意味を持つ異なる言語で書かれた文ペアの集合から確率的翻訳規則を自動学習し、それをもとに入力文に対する各翻訳候補を順位付け、最も確率の高い候補を出力とする¹⁾。フレーズベース翻訳は、連続する1単語以上の単語列（フレーズ）を最小単位として翻訳を行う手法であり^{2),3)}、近年、統計的機械翻訳の主流となっている。しかし、入力文に対するフレーズ区切りや翻訳前後のフレーズどうしの対応（フレーズ対応）の組合せ数が膨大であるため、フレーズベースシステムは大胆な近似や様々なヒューリスティクスを用いて探索空間を狭めている²⁾⁻⁵⁾。したがって、一般にフレーズベースシステムの出力は確率最大の翻訳候補とは限らない。近似計算の結果、確率値が第2位以下となり翻訳結果として採択されなかった候補の中には、より適切な翻訳文が含まれることが知られており⁶⁾、確率を厳密に最大化するフレーズ区切り・対応を各翻訳候補に対して求め直し、その順位付けをより正確にすることで、翻訳精度が改善されると考えられる。

^{†1} 筑波大学大学院システム情報工学研究科

Graduate School of Systems and Information Engineering, University of Tsukuba

^{†2} 情報通信研究機構 MASTAR プロジェクト

MASTAR Project, National Institute of Information and Communications Technology

^{†3} 大阪大学大学院情報科学研究科

Graduate School of Information Science and Technology, Osaka University

^{†4} 中央大学理工学部情報工学科

Department of Information and System Engineering, Faculty of Science and Engineering, Chuo University

対訳文に対して、与えられたフレーズ翻訳規則の集合から確率最大すなわち最適なフレーズ対応 (phrase alignment) を求める手法としては、整数線形計画法を用いた手法⁷⁾ が提案されている。しかし、この手法では、日本語-英語のような語順変化の激しい言語間の翻訳において、精度の向上に大きく寄与する語順並べ替え (リオーダリング) モデル⁴⁾ が考慮されていない。本稿では、リオーダリングモデルを組み込み可能な新しいフレーズアライナ (phrase aligner) の定式化を提案し、これを統計的機械翻訳システムの翻訳候補のリランキングに応用することで翻訳精度が向上することを実験的に示す。

以下、2章において統計的機械翻訳の概要や従来手法の問題点を説明し、3章で提案手法について解説する。その後、4章で評価実験について報告し、5章で本稿のまとめを行う。

2. フレーズアライナ

2.1 統計的機械翻訳

統計的機械翻訳システムは、入力された原言語文 f に対する翻訳候補としてあらゆる目的言語文 e を考慮し、最も確率の高い目的言語文 \hat{e} を翻訳結果として出力する。ここで、原言語および目的言語はそれぞれ翻訳元の言語、翻訳先の言語を意味する。統計的機械翻訳では、以下の近似式に基づいて出力文 \hat{e} が決定される¹⁾。

$$\begin{aligned} \hat{e} &= \arg \max_e P(e) \sum_c P(f, c|e) \\ &\approx \arg \max_e P(e) \max_c P(f, c|e) \\ &\approx \arg \max_{e,c} P(e) P(f, c|e) \end{aligned} \quad (1)$$

式 (1) において、 $P(e)$ は言語モデルと呼ばれ、 e の目的言語文らしさを表す。また、 c は f, e の対訳関係を表す。式 (1) の 2 段目では、最も尤もらしいフレーズ対応に確率が集中するという仮定のもと、 \sum_c を \max_c で近似している。なお、この近似は一例であり、実際には様々な近似の方法が存在する。 $\arg_e \max_{e,c}$ はデコーダと呼ばれ、入力文 f の翻訳として最も確率の高い e をヒューリスティックに探索する。ここで \max' はヒューリスティック探索による \max の近似を表す。ヒューリスティック探索としては、現在、マルチスタック・ビームサーチ法²⁾ が最も広く用いられている。

実際には、式 (1) を用いるのではなく、式 (2) のような対数線形モデルにより各素性の統合を行う³⁾。

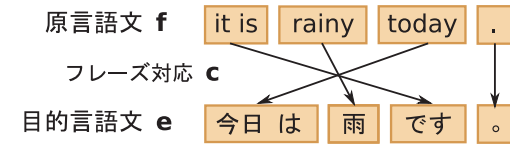


図 1 フレーズベース翻訳の例

Fig. 1 An example of phrase-based translation.

$$\hat{e} = \arg \max_e \sum_{c} \lambda_k h_k(f, c, e) \quad (2)$$

ここで、 $h_k(f, c, e)$ は k 番目の素性を表し、 λ_k はその重みである。対数線形モデルの枠組みでは、式 (1) に表れた確率モデルを対数化したもののほか、式 (1) から導くことのできない素性 (たとえば、目的言語文の長さに対するペナルティなど) も翻訳精度向上のために利用される。なお本稿では、各素性値のことをスコアと呼ぶ。

2.2 フレーズベース統計的機械翻訳

フレーズベース手法では、連続した 1 単語以上の単語列であるフレーズを翻訳の最小単位とする²⁾。フレーズベース翻訳の例を図 1 に示す。枠で囲われた単語ないしは単語列がフレーズを、原言語フレーズと目的言語フレーズとを結ぶ矢印がフレーズ対応を表しており、矢印の両端の 2 フレーズをフレーズ対と呼ぶ。フレーズベース翻訳では以下の近似をおく²⁾。

$$\begin{aligned} P(f, c|e) &= P(c|e) P(f|c, e) \\ &\approx P(c_1^I|e) \prod_{i=1}^I P(\bar{f}_{c_i}|\bar{e}_i) \end{aligned} \quad (3)$$

ここで、 \bar{f} 、 \bar{e} はそれぞれ原言語フレーズ、目的言語フレーズを、 $c_1^I = c_1, c_2, \dots, c_I$ はフレーズ対応を表す。 c_i は目的言語側で i 番目のフレーズ \bar{e}_i が対応する原言語側フレーズの番号である。すなわち、 \bar{e}_i は原言語側で c_i 番目のフレーズ \bar{f}_{c_i} に対応する。 $P(\bar{f}_{c_i}|\bar{e}_i)$ はフレーズ翻訳確率である。 $P(c_1^I|e)$ はリオーダリングモデルと呼ばれ、翻訳前後での語順変化に対する確率を与える。

2.3 リオーダリングモデル

フレーズベース手法におけるリオーダリングに関する素性としては、Lexicalized Block Orientation モデル⁴⁾ および歪みペナルティ²⁾ が広く用いられている。どちらも目的言語側

で i 番目および $(i+1)$ 番目すなわち目的言語側が隣接しているフレーズ対のペアに対する局所的なりオーダリングの評価を行う。本節では、これらのモデルについて説明する。

2.3.1 Lexicalized Block Orientation モデル

Lexicalized Block Orientation (LBO) モデルでは、目的言語側で i 番目および $(i+1)$ 番目のフレーズ \bar{e}_i, \bar{e}_{i+1} に対応するフレーズ対のペアに対して、その原言語側フレーズ \bar{f}_{c_i} と $\bar{f}_{c_{i+1}}$ との位置関係を以下のような 3 クラスに分類する⁴⁾。

$$\text{class}(c_i, c_{i+1}) = \begin{cases} \text{monotone} & (c_{i+1} = c_i + 1 \text{ のとき}), \\ \text{swap} & (c_{i+1} = c_i - 1 \text{ のとき}), \\ \text{discontinuous} & (\text{それ以外}). \end{cases} \quad (4)$$

目的言語側で隣接するフレーズ対に対して、*monotone* は原言語側において隣接かつ目的言語側と同順であること、*swap* は原言語側において隣接かつ目的言語側と逆順であること、*discontinuous* は原言語側において両者が離れていることを表す⁴⁾。たとえば図 1 において、“今日は”–“today” のフレーズ対と “rainy”–“雨” のフレーズ対とは *swap* の関係をなしている。

LBO モデルでは、上記の 3 クラスを用いた各フレーズ対間のリオーダリング確率の積として、式 (3) 第 1 項を以下のように近似する⁴⁾。

$$P(c_1^I | e) \approx \prod_{i=1}^I P(\text{class}(c_i, c_{i+1}) | \bar{e}_i, \bar{e}_{i+1}) \quad (5)$$

対数線形モデルの枠組みにおいては翻訳精度向上のために、 $P(\text{class}(c_i, c_{i+1}) | \bar{e}_i, \bar{e}_{i+1})$ の代わりとして、より多くの情報を持つ $P(\text{class}(c_i, c_{i+1}) | \bar{f}_{c_i}, \bar{f}_{c_{i+1}}, \bar{e}_i, \bar{e}_{i+1})$ を素性を使用する場合が多い。

2.3.2 歪みペナルティ

歪みペナルティは、目的言語側が隣接しているフレーズ対の原言語側フレーズ \bar{f}_{c_i} と $\bar{f}_{c_{i+1}}$ との間に含まれる単語数 (単語距離) に比例してリオーダリングに対するペナルティを与え²⁾。すなわち、この素性は翻訳結果の *monotone* 性に対してバイアスをかける。文頭を 0 としたときの \bar{f}_{c_i} の終端単語位置を end_i 、 $\bar{f}_{c_{i+1}}$ の開始単語位置を $start_{i+1}$ とすれば、歪みペナルティ $d(end_i, start_{i+1})$ は以下のように表される²⁾。

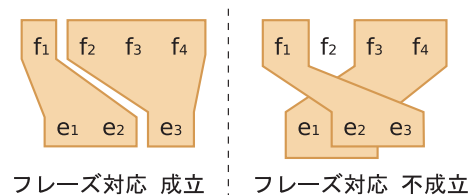


図 2 フレーズ対応の成立例 (左図) と不成立例 (右図)

Fig. 2 Examples of correct phrase alignment (left) and incorrect phrase alignment (right).

$$d(end_i, start_{i+1}) = |end_i - start_{i+1} + 1| \quad (6)$$

なお、歪みペナルティは式 (3) から導かれるモデルではなく、対数線形モデルの枠組みでのみ考慮可能な素性である。

2.4 フレーズ対応問題

対訳文の各単語を 1 度ずつ被覆するフレーズ対の組合せをその対訳文に対する全単射フレーズ対応という⁷⁾。なお、本稿においてはフレーズ対応として全単射フレーズ対応のみを扱うので、以下では単にフレーズ対応と表記する。

フレーズ対応が成立する場合とそうでない場合の例を図 2 に示す。図中、 $f_1 \sim f_4$ と $e_1 \sim e_3$ は、それぞれ原言語文、目的言語文を構成する単語列を表し、両単語列は対訳文をなしている。また、枠内はフレーズ対である。図 2 左図は、フレーズ対群により 1 度ずつ各単語が被覆されるため、フレーズ対応が成立している。一方、図 2 右図には、被覆されない単語や重複して被覆される単語が含まれるため、フレーズ対応は不成立となる。

フレーズ対応問題とは、対訳文 $\langle f, e \rangle$ 、およびフレーズ対とその翻訳確率 $\langle \bar{f}_{c_i}, \bar{e}_i, P(\bar{f}_{c_i} | \bar{e}_i) \rangle$ の集合であるフレーズテーブルが与えられたとき、確率最大すなわち最適なフレーズ対応を求める問題であり⁷⁾、これを実現するシステムを本稿においてはフレーズアライナと呼ぶ。フレーズアライナは以下の式で定義される。

$$\langle \hat{f}_1^I, \hat{e}_1^I, \hat{c}_1^I \rangle = \arg \max_{I, c_1^I, \bar{f}_1^I = f, \bar{e}_1^I = e} P(\bar{f}_1^I | \bar{e}_1^I, c_1^I) P(\bar{e}_1^I, c_1^I | e) \quad (7)$$

式 (7) の右辺第 1 項および第 2 項はそれぞれフレーズ翻訳確率、リオーダリング確率を表し、式 (3) の第 2 項、第 1 項と対応している。なお、フレーズ対応問題はフレーズベース翻訳の部分問題であり、フレーズアライナは式 (1) 2 段目の \max_c 以降に相当する。

フレーズ対応問題において、フレーズ長に制限をおかない場合、フレーズ対応になりうるフレーズ対の組合せ、なりえないフレーズ対の組合せを合わせて $\sum_{I=1}^{\min(|f|,|e|)} \binom{\xi}{I}$ 個の候補が存在する。ここで、 I はフレーズ対応に使用されるフレーズ対の数、 ξ は対訳文に適用可能なフレーズ対数である。本稿の実験で使用したデータにおいては、 ξ の平均は 500 程度、 $\min(|f|,|e|)$ は平均で 30 程度であった。これらの平均数を用いると、フレーズ対応問題の解候補数は、およそ 10^{48} 個にもものぼる。

2.5 従来手法による整数線形計画問題としての定式化

フレーズ対応問題は非常に大規模な問題であり、解を全列挙することは現実的には不可能である。そこで、フレーズ対応問題を整数線形計画問題として定式化し、一般的な数理計画問題の solver を利用して最適フレーズ対応を求める手法⁷⁾ が提案されている。その説明を以下で行う。

フレーズテーブル中のフレーズ対 $\langle \bar{f}_{c_k}, \bar{e}_k \rangle$ に対して、その使用の有無を表す 2 値変数 $x_k \in \{0, 1\}$ 、および各フレーズ対が原言語文で被覆する単語位置を 1、それ以外を 0 で表す 2 値行列 F を定義する⁷⁾。 F は行が原言語文に含まれる各単語、列が各フレーズ対に対応している。目的言語側についても F と同様に行列 E を定義する⁷⁾。

例として、対訳文 $\langle f = f_1, f_2, f_3, f_4, e = e_1, e_2, e_3 \rangle$ に対し、図 3 のようなフレーズ対が適用可能な場合を考える。図 3 において枠内がフレーズであり、フレーズどうしを結ぶ線がフレーズ対応を表す。このとき、 F, E は式 (8) のようになる。たとえば、フレーズ対 1 は原言語側で f_1, f_2 を被覆するため、行列 F の 1 行 1 列および 2 行 1 列成分が 1 となっている。

$$F = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, E = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (8)$$

従来手法では簡化のために、式 (7) 右辺第 2 項のリオーダーリング確率を定数とおき、翻訳確率のみを考慮する⁷⁾。この仮定のもとでは、フレーズ対応問題は次のように表される⁷⁾。

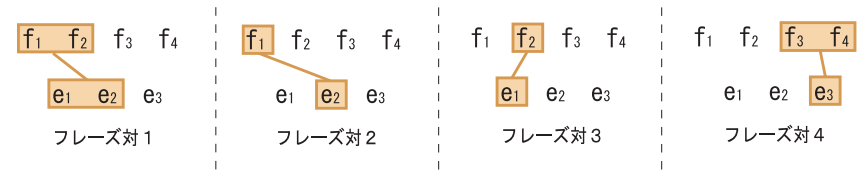


図 3 対訳文に対して適用可能なフレーズ対集合
Fig. 3 Candidates of phrase pairs for input parallel sentences.

$$\begin{cases} \text{maximize} & \sum_{k \in K} x_k t_k \\ \text{subject to} & Fx = \mathbf{1}, \\ & Ex = \mathbf{1}, \\ & x_k \in \{0, 1\} \quad (\forall k \in K). \end{cases} \quad (9)$$

ここで t_k は、 $\lambda_{tm} \log P(\bar{f}_{c_k} | \bar{e}_k)$ の略記であり、 λ_{tm} は翻訳確率に対する重みである。 K は対訳文に適用可能なフレーズ対の集合を表す。また $\mathbf{1} = (1, \dots, 1)^T$ 、 $x = (x_1, \dots, x_K)^T$ である。制約式 $Fx = \mathbf{1}$ において、 Fx は使用するフレーズ対群によって被覆される原言語単語を表し、右辺の 1 は、対訳文中のすべての原言語単語が 1 度ずつ被覆されることを表す。目的言語側制約条件 $Ex = \mathbf{1}$ についても同様である。目的関数では、フレーズ対応に使用されているフレーズ対の翻訳確率について最大化を行っている。

3. 提案手法

3.1 フレーズアライナを用いたリランキン

本稿では、フレーズアライナを利用した翻訳候補のリランキン法を提案する。リランキンの流れを図 4 に示す。まず、フレーズベースデコーダを用いて入力文に対する翻訳候補上位 N 個すなわち N-best 翻訳を求める。次に、デコーダに入力した原言語文と各翻訳候補のペアを対訳文として、フレーズアライナに入力し、その最適フレーズ対応を求める。なお、翻訳時と異なりフレーズ対応最適化時には、目的言語文が固定されているため、言語モデルや目的言語文長ペナルティのような目的言語文のみに依存して決まる素性値は考慮しなくてもよいことに注意されたい。最後に、最適化されたフレーズ対応のスコアに対して、言語モデルスコアおよび目的言語文長ペナルティを加味し、各翻訳候補のリランキンを行い、最も確率の高い候補を翻訳システムの出力とする。すなわち、提案手法は、従来のフレーズベースデコーダ (式 (1) の $\arg_e \max'_{e,c}$) とは異なり、 c についての最大化のみでは

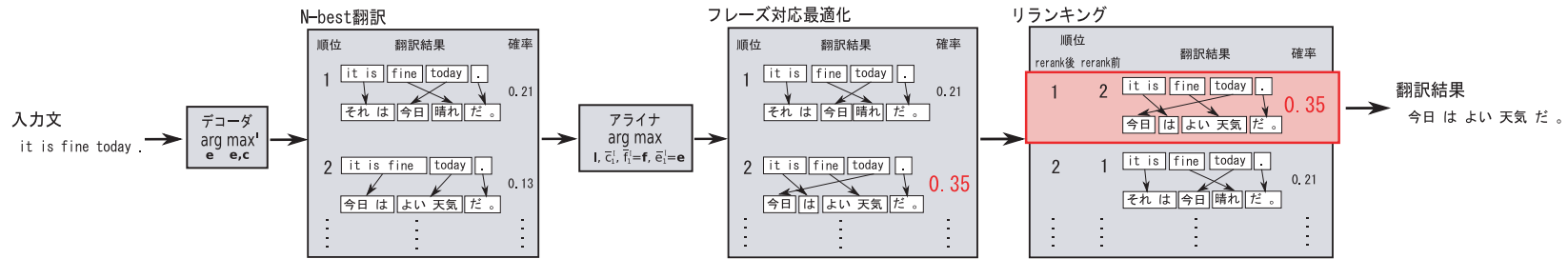


図 4 フレーズアライナによる翻訳候補のリランキングの流れ
Fig. 4 Flowchart of N-best reranking method using phrase aligner.

あるが、近似探索 \max'_e ではなく厳密に \max_e として行う。

3.2 フレーズ対応問題の新しい定式化

2.5 節で説明した従来手法によるフレーズ対応問題の定式化では、各フレーズ対に対して変数をおくため、フレーズ対どうしの位置関係を表すリオーダーリングモデルを 1 次式として目的関数に組込むのは難しい。LBO モデルや歪みペナルティは 2 つのフレーズ対間の関係を表すため、従来法ではリオーダーリングスコアの表現に 2 つの変数を必要とする。たとえば、フレーズ対 2 とフレーズ対 3 とのリオーダーリングスコアは、両フレーズ対が対応の中で使用されているときのみ有効となることを表すために x_2 と x_3 の積が必要となる。

しかし、一般に 2 次以上の非線形計画問題は、線形計画問題と比較すると効率的に探索を行うことができない。そこで本稿では、1 次式のみによりリオーダーリングスコアを表現するフレーズ対応問題の新たな定式化を提案する。

本稿で提案する定式化では、対訳文に適用可能なフレーズ対群を目的言語側についてのみ有向グラフ上に表現することを考える。図 3 に示された各フレーズ対の目的言語フレーズを有向グラフとして表すと図 5 のようになる。図 5 において $e_1 \sim e_3$ は目的言語文中の単語である。フレーズは枠で表されており、その番号は図 3 のフレーズ対の番号と対応している。また、各フレーズの位置と大きさは被覆する単語位置に対応している。フレーズどうしを結ぶ有向枝に割り振られた数字は枝番号を表す。フレーズ対応は、従来法と同様に原言語側について式 (9) の原言語側単語被覆制約 $F\mathbf{x} = \mathbf{1}$ を満たし、かつ目的言語側グラフにおいて開始 (文頭) ノード s から終端 (文末) ノード g へのパスとなっている必要がある。

たとえば、図 5 においてフレーズ 3, 2, 4 を通るパスは、原言語側制約も同時に満たしているため (図 3 参照)、対訳文に対するフレーズ対応候補の 1 つである。LBO モデルお

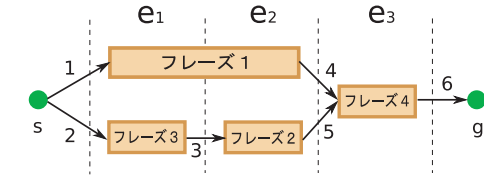


図 5 有向グラフ上に表した図 3 の目的言語フレーズ群
Fig. 5 Directed graph built from the target side phrases in 図 3.

よび歪みペナルティは、目的言語側で隣接するフレーズ対ペアを評価するモデルであるので、提案する定式化のもとでは、リオーダーリングスコア (LBO モデルスコアと歪みペナルティの合計値) を目的言語側グラフの枝に対する重みとして組み込むことができる。

式 (9) の定式化において使用した記号に加えて、目的言語側有向グラフ中の枝 a に対し、 a がパスに含まれる場合は 1, そうでない場合は 0 をとる仮変数 y_a を新たに定義する。このとき、フレーズ対応問題は以下のように定式化される。

$$\begin{aligned}
 & \text{maximize} && \sum_{k \in K} x_k t_k + \sum_{a \in A} y_a r_a \\
 & \text{subject to} && F\mathbf{x} = \mathbf{1}, \\
 & && M\mathbf{y} = \mathbf{b}, \\
 & && N\mathbf{y} = \mathbf{x}, \\
 & && x_k \in \{0, 1\} \quad (\forall k \in K), \\
 & && y_a \in \{0, 1\} \quad (\forall a \in A).
 \end{aligned} \tag{10}$$

$My = b$ は目的言語側グラフで文頭ノード s から文末ノード g へのパスとなるための条件であり流量保存則と呼ばれる。 b の各要素は、ノードへの流入を $+1$ 、ノードからの流出を -1 と表現する場合、 $b_s = -1$ 、 $b_g = +1$ 、 $b_{others} = 0$ となる。ここで、 b_s は文頭ノード s 、 b_g は文末ノード g 、 b_{others} はそれ以外の各ノードに対応する。それぞれ、グラフの始点からは必ず流出のみ、グラフの終点には必ず流入のみが生じ、その他のノードでは流入量と流出量が一致することを表している。

N は目的言語側パスに含まれる枝と各フレーズ対との関係を与える行列であり、 $Ny = x$ は仮変数 y から変数 x へ変換を行う式である。以下、式 (10) において $Ny = x$ が必要となる理由について説明する。

例として、図 3 においてフレーズ対 1 と 4 を使う場合、図 5 において枝 2, 3, 5, 6 を通るパスを考える。前者は原言語側制約 $Fx = 1$ を満たすことは明らかであり、後者は目的言語側制約 $My = b$ を満たし、そのパスにフレーズ対 3, 2, 4 のノードを含む。 $Fx = 1$ 、 $My = b$ は、それぞれ変数 x 、 y にのみ依存するため、 $Ny = x$ が式 (10) に含まれない場合、上記のような原言語側と目的言語側で用いられているフレーズ対が異なる組合せも解に含まれてしまう。これを防ぐために、変数 x と y とに一意的な関係を持たせる目的で変数変換式 $Ny = x$ が必要となる。

また A はすべての枝の集合を、 r_a は目的言語側の枝 a に割り当てられるリオーダーリングスコアを表す。なお、 r_a の値はリオーダーリング素性値とその重みの積である。目的関数第 1 項、第 2 項はそれぞれフレーズ翻訳スコア、リオーダーリングスコアに対応しており、従来手法では 1 次式として表現できなかったリオーダーリングスコアを、提案手法では第 2 項により表すことが可能となっている。

3.3 流量保存則、変数変換式の具体例

図 5 について、流量保存則 $My = b$ を書き下すと式 (11) のようになる。左辺第 1 項は行列 M であり、その各行は s 、フレーズ 1、 \dots 、4、 g のノードに、各列は枝に対応する。たとえば、式 (11) の 5 行目 $y_4 + y_5 - y_6 = 0$ は、フレーズ対のノードについての流量保存則に対応し、フレーズ 4 のノードへの流入量 $y_4 + y_5$ が流出量 y_6 と等しいことを表す。

$$\begin{pmatrix} -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad (11)$$

また、図 5 のグラフにおいてフレーズ 4 のノードには枝 4 ないしは枝 5 からパスが流入することになる。したがって枝に対する変数 y_4 、 y_5 のいずれかが 1 をとるとき、フレーズ対 4 に対する変数 x_4 の値は 1 となる。以上より $y_4 + y_5 = x_4$ という関係式が導かれる。その他のフレーズ対についても同様に有向枝との関係を求めると、仮変数 y から変数 x への変数変換式 $Ny = x$ は次のようになる。

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad (12)$$

ここで、左辺第 1 項の行列が、仮変数 y から変数 x を導出するための変換行列 N である。

4. 評価実験

4.1 実験条件

評価実験には、NTCIR-7 特許文翻訳タスク⁸⁾ で配布された英日対訳コーパスを用いた。フレーズテーブルおよびリオーダーリングモデルの学習には同コーパス学習セットを使用した。デコーダの各素性に対する重み学習は、development セットを用いて Minimum Error-Rate Training⁹⁾ により行った。また、テストセットは NTCIR-7 フォーマルランで配布されたテストセットとした。コーパスの詳細を表 1 に示す。

翻訳方向は日本語から英語とし、翻訳精度の評価には BLEU¹⁰⁾ を、ベースラインには

表 1 NTCIR-7 特許文翻訳タスクコーパス詳細

Table 1 Description of the data set on NTCIR-7 patent translation task.

データセット	文数	単語数	語彙サイズ
学習セット (英)	1,798,571	59,974,173	142,435
学習セット (日)	1,798,571	64,184,179	121,652
dev. セット (英)	915	30,028	3,986
dev. セット (日)	915	32,427	3,653
テストセット (英)	1,381	45,334	4,116
テストセット (日)	1,381	48,737	3,882

表 2 評価実験条件

Table 2 Experimental condition.

項目	条件
デコーダ	Moses (08/02/20 release)
ビーム幅 (翻訳候補数)	10, 20, 50, 100, 200, 500, 1,000
ttable-limit	20
リオーダーリングモデル	msd-bidirectional-fe
歪み距離制限	なし
言語モデル	5gram, Interpolated Modified Kneser-Ney
整数計画問題の solver	ILOG CPLEX 11.0

Moses デコーダ⁵⁾を用いた^{*1}。また N-best 翻訳を得る際、目的言語文のバリエーションを増やすため、フレーズ区切りは異なるが同一の目的言語文を導く翻訳候補群は、その中でスコア最大の候補のみを N-best に含めるようにした。言語モデルの学習は、学習セットの英語側を用いて SRI Language Modeling Toolkit¹¹⁾により行った。デコーダパラメータなどの実験条件を表 2 に示す。ビーム幅は Moses デコーダの探索精度を与えるパラメータであり、原言語文 1 文あたりの N-best 数と同値とした。整数計画問題の solver には、CPLEX 11.0¹²⁾を用いた。

4.2 実験結果

ベースラインおよび提案手法による翻訳結果の BLEU を図 6 に示す。図中、Moses はベースラインを、rerank は提案手法をそれぞれ意味する。提案手法と Moses とを比較すると、提案手法は常に Moses より BLEU 値が高くなっている。有意水準 5% として検定を

*1 Moses の N-best 翻訳のリオーダーリングスコアには近似が含まれている。本研究ではスコア付けの厳密化を目的とするため、このような近似を行わないよう Moses の一部を実装しなおした。

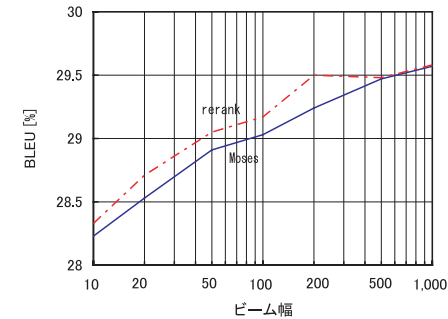


図 6 ビーム幅に対する翻訳精度

Fig. 6 Effect of the beam width on translation quality (BLEU).

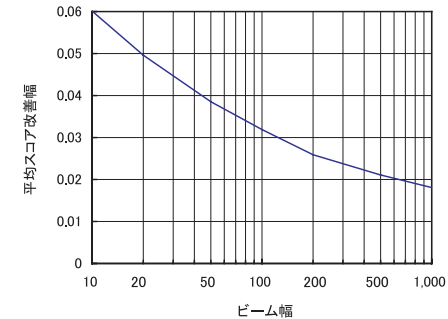


図 7 ビーム幅に対する平均スコア改善幅

Fig. 7 Effect of the beam width on the average score improvement.

行ったところ、ビーム幅 200 以下の全測定点において、提案手法が Moses を有意に改善するという結果が得られた。一方、ビーム幅 500 以上では Moses と提案手法とで翻訳精度はほぼ変わらない。これはビーム幅を大きくとることで、Moses の探索の精度が向上するためである。

従来のヒューリスティック探索の精度を調べるため、N-best 各候補のフレーズ対応最適化前後のスコアの差分の平均 (平均スコア改善幅) とビーム幅との関係性を求めた結果を図 7 に示す。平均スコア改善幅が大きいほど、従来法でのフレーズ対応に関する探索精度が悪いことを表す。図 7 では、ビーム幅を大きくしていくと平均スコア改善幅が小さくなっていく様子が見られる。この結果は、図 6 においてビーム幅が大きいとき BLEU 改善が小さかった

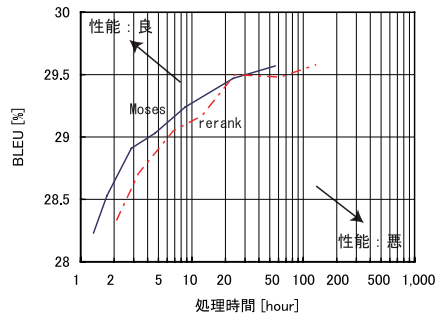


図 8 翻訳時間と BLEU との関係

Fig. 8 Effect of the search time on translation quality (BLEU).

ことに対する裏づけとなっている。また、各候補のスコアの大きさはおよそ $-10 \sim -100$ 程度のオーダーであり、スコアの改善幅はこれにくらべて数千分の 1 の大きさである。したがって、従来法でもフレーズ対応については十分な探索精度が得られているといえる。

テストセットに対する処理時間と BLEU の関係を図 8 に示す。ここで処理時間は、Moses の場合にはテストセット全文の翻訳に要した時間を、rerank の場合には Moses の翻訳時間とフレーズ対応最適化に要した時間との和を表す。Moses と提案手法とで処理時間を比較すると、翻訳精度が同じであるとき、提案手法はおよそ Moses の 2 倍程度の時間を必要とすることが分かる。

5. おわりに

本稿では、リオーダーリングモデルを考慮したフレーズ対応問題の新しい定式化を提案した。また、提案した定式化を用いてフレーズベース翻訳システムの N-best 翻訳に対し、フレーズ対応のスコアを厳密に最大化することによってリランキングを行う手法について検討を行った。

NTCIR-7 特許文翻訳タスクのデータセットを用いた評価実験において、提案手法は Moses が翻訳候補に与えたスコアをつねに改善し、BLEU を有意に向上させることが確認できた。一方で、提案手法ではすべてのフレーズ対応候補の中から最適なフレーズ対応を探索しているにもかかわらず、スコアの改善幅はそれほど大きくないという結果が得られた。提案手法は、フレーズ対応付き翻訳候補文のスコアの厳密値を求めている。そのため、従来のヒューリスティック探索であっても、フレーズ対応については高い精度の解が得られていることが

明らかになった。

デコーダの探索空間の大きさはフレーズ対応と目的言語文のバリエーションによって決まるが、従来法でもフレーズ対応についての探索精度は十分であることから、目的言語文についての探索精度に問題があると考えられる。したがって、フレーズベース翻訳における探索の課題は、より多くの目的言語文のバリエーションを考慮することにあるという示唆が得られる。

この示唆をふまえ、今後は、リランキング法のごく限られた数の翻訳候補に対して最適化手法を利用するのではなく、デコーダ内部で利用することで、より多くの翻訳候補についてスコアの最適化を行う方法について検討していきたい。

参考文献

- 1) Brown, P.F., Pietra, V.J.D., Pietra, S.A.D. and Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol.19, No.2, pp.263-311 (1993).
- 2) Koehn, P., Och, F.J. and Marcu, D.: Statistical Phrase-Based Translation, *Proc. Human Language Technology and North American Association for Computational Linguistics Conference* (2003).
- 3) Och, F.J. and Ney, H.: The Alignment Template Approach to Statistical Machine Translation, *Computational Linguistics*, Vol.30, No.4, pp.417-449 (2004).
- 4) Tillmann, C. and Zhang, T.: A Localized Prediction Model for Statistical Machine Translation, *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.557-564 (2005).
- 5) Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Association for Computational Linguistics, pp.177-180 (2007).
- 6) Hasan, S., Zens, R. and Ney, H.: Are Very Large N-Best Lists Useful for SMT?, *Proc. Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, Association for Computational Linguistics, pp.57-60 (2007).
- 7) DeNero, J. and Klein, D.: The Complexity of Phrase Alignment Problems, *Proc. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology, Short Papers*, Association for Computational Linguistics, pp.25-28 (2008).

- 8) Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro, T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, NTCIR, pp.389–400 (2008).
- 9) Och, F.J.: Minimum Error Rate Training in Statistical Machine Translation, *Proc. 41st Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp.160–167 (2003).
- 10) Papineni, K., Roukos, S., Ward, T. and Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.311–318 (2002).
- 11) Stolcke, A.: SRILM – An Extensible Language Modeling Toolkit, *Proc. International Conference on Spoken Language Processing*, pp.901–904 (2002).
- 12) ILOG: *ILOG CPLEX 11.0 User's Manual*, ILOG (2007).

(平成 21 年 11 月 10 日受付)
(平成 22 年 5 月 6 日採録)



越川 満 (正会員)

1986 年生。2010 年筑波大学大学院システム情報工学研究科博士前期課程修了。同年株式会社バンダイナムコゲームス入社。現在、同社未来研究所に勤務。大学院在学中は、自然言語処理、特にフレーズベース統計的機械翻訳の研究に従事。



内山 将夫 (正会員)

1969 年生。1992 年筑波大学卒業。1997 年同大学院工学研究科修了。博士 (工学)。現在、情報通信研究機構主任研究員。言語処理の実際的で学際的な応用に興味がある。言語処理学会、ACL 等各会員。



梅谷 俊治 (正会員)

1974 年生。1998 年大阪大学大学院基礎工学研究科情報数理系専攻博士前期課程修了。2004 年京都大学大学院情報学研究所数理工学専攻博士後期課程指導認定退学。2008 年より大阪大学大学院情報科学研究科准教授。博士 (情報学)。組合せ最適化、アルゴリズムに関する研究に従事。日本 OR 学会、INFORMS、MPS 各会員。



松井 知己

1990 年 3 月東京工業大学大学院総合理工学研究科システム科学専攻博士後期課程修了。同年 4 月東京理科大学理工学部経営工学科助手。1992 年 4 月東京大学工学部計数工学科講師。1996 年 4 月東京大学大学院工学系研究科計数工学専攻助教授。2006 年 4 月中央大学理工学部情報工学科教授となり現在に至る。組合せ最適化の研究に従事。博士 (理学)。オペレーションズ・リサーチ学会等の会員



山本 幹雄 (正会員)

1961 年生。1986 年豊橋技術科学大学大学院修士課程修了。同年 (株) 沖テックノシステムズラボラトリ研究開発員。1988 年豊橋技術科学大学情報工学系教務職員。1995 年筑波大学電子・情報工学系講師。2008 年筑波大学システム情報工学研究科教授。博士 (工学)。自然言語処理の研究に従事。言語処理学会、人工知能学会、ACL 各会員。