複数の目的言語の同時生成による統計的機械翻訳

Graham Neubig, Philip Arthur, Kevin Duh
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

通常の統計的機械翻訳 (statistical machine translation; SMT) システムでは、原言語文 F から 1 つの目的言語文 E へと翻訳を行う。しかし、人手翻訳が実際に行われる場面では、同一の内容を多数の言語へ翻訳することが多い。例えば、国連文章は 6 カ国語へ翻訳され [5]、TED 講演の字幕は 50 カ国語以上へ翻訳されている [2]。しかし、機械翻訳の研究では、このような多言語データを有効に利用する手法は存在しながらも比較的少ない。複数の原言語文の翻訳結果を組み合わせる研究 [10] や多くの言語でトピックモデルを学習する研究 [7] はその例として挙げられる。

本研究では、多言語データを有効に利用する手法として、**複数の目的言語文**を同時生成する機械翻訳手法を提案する。つまり、F から単一の E を生成するのではなく、複数の目的言語からなる文の集合 $\langle E_1, E_2, \ldots, E_N \rangle$ を生成する（以降、各目的言語文に対応する言語を T1、T2 などと省略する）。この枠組みは、英語の国連文章や TED 講演から複数の言語の文章や字幕を作成することに類似しているとも言える。

しかし、複数の目的言語文を同時生成するにはどのようなメリットがあるのか？アラビアを原言語、英語を T1、日本語を T2 とした翻訳例を図 1 に示す。日本語母語話者が英語の文だけを閲覧した時に、正しい訳を選択できるのだろうか？英語の能力が高い、もしくは政治に詳しい一部の人を除けば、悩む人は多かろう。しかし、英語能力が高くなくても、対応する T2 の日本語訳も与えられれば、容易に 3 つ目の訳を最も自然として選択できる可能性が高い。

この例を SMT の枠組みに言い換えると、E_1 と E_2 という 2 つの目的言語文を同時に生成し、E_1 を選択する際に E_2 の自然性を判断材料に利用することに当たる。語句の自然性を判定する役割を担うのは言語モデル (language model; LM) であり、大量で良質なデータを使うほど高性能な LM ができ、翻訳精度が向上する [1]。つまり、T1 では小さな LM、T2 では大きな LM が構築できる場合、E_2 の自然性が E_1 の選択の助けになる場合は容易に考えられる。例えば、言語資源の少ない言語や、新しく国連に加入した国の言語などが具体例として思い浮かぶ。

複数の目的言語を扱う SMT の具体的な実現方法として、Chiang[4] の同期文脈自由文法 (synchronous context free grammar; SCFG) を採用する (2 節)。SCFG は 2 言語の文 F と E を同時生成するが、我々はこの枠組みを任意の数の文 $\langle F, E_1, E_2, \ldots, E_N \rangle$ を同時生成できる複数同期文脈自由文法 (multi-synchronous CFG; MSCFG) へとさらに拡張する (3 節)。MSCFG を対訳データから学習する方法 (4 節) と、複数の目的言語に対する LM を考慮する探索方法 (5 節) を提案する。提案法の有効性を検証するために、国連文章を題材に実験的評価を行う (6 節)。その結果、T1 に比べて T2 で高性能な LM が構築でき、T1 と T2 が比較的類似している言語である場合において、複数の目的言語を考慮することで性能の向上が実現可能であることが分かった。

図 1: 複数の目標言語を考慮した翻訳の例

2 同期文脈自由文法

まず、階層的フレーズベース翻訳 (Hiero [4]) を初めとする様々な翻訳方式で利用される SCFG について紹介する。SCFG は X、γ、α の 3 つ組で表現する同期ルールからなる。

$$X \rightarrow \langle \gamma, \alpha \rangle \tag{1}$$

ここで、X はルールの親記号であり、γ と α はそれぞれ原言語と目的言語における終端記号と非終端記号からなる記号列である。非終端記号はすべてインデックスを持っており、γ と α で同じインデックスを持つ記号は対応すると見なされる。同期ルールの例を下に示す。

$$X \rightarrow \langle X_0 \text{ of the } X_1,\ X_1 \text{ の } X_0 \rangle. \tag{2}$$

原言語文 F が与えられた際、文に合致するルールを探し出し、CKY+アルゴリズム [3] で構文解析を行うと生成可能な翻訳結果が列挙でき、さらにルールにスコアが付与された場合の SCFG であれば、スコア最大の翻訳仮説も取得可能である。また、目的言語文に対する LM を考慮する際、計算量は増加するが、キューブ枝狩り [4] などの近似法を導入することで現実的な時間で訳の生成が可能となる。

3 複数同期文脈自由文法

本節では、複数の目的言語文の同時生成を可能とする提案法の MSCFG について述べる。書き換えルールが原言語記号列 γ と目的言語記号列 α からなる通常の SCFG を拡張し、下記の通り N 個の目的言語記号列を有するルールを定義する。

$$X \rightarrow \langle \gamma, \alpha_1, ..., \alpha_N \rangle \tag{3}$$

MSCFG を用いて訳出を行うには、通常の SCFG と同じく、CKY+アルゴリズムで原言語文の構文解析を行う。構文解析の結果得られる導出から目的言語文を生成する際、1 つだけではなく、複数の目的言語文を生成する。

この定式化は通常の SCFG とさほど変わらない単純な拡張であるとも言える。しかし、翻訳システムの学習と訳出を行う上で、複数の目的言語文に対応する新たなアルゴリズムが必要となり、これらを以降の節で提案する。

4 MSCFG の学習

本節では、複数の言語の対訳コーパスから、MSCFG に基づく翻訳モデルを学習する手法について述べる。

4.1 ルール抽出

まず、Chiang[4] による通常の 2 言語 SCFG のルール抽出について簡単に述べる。原言語コーパス $\mathcal{F}$ と目的言語コーパス $\mathcal{E}$ を準備し、各文に対して IBM モデルなどを用いて単語アライメントを得る [11]。

この文とアライメントに基づき、フレーズ抽出を行う。原言語文 $F = f_1^J$ と目的言語文 $E = e_1^I$、原言語インデックス i と目的言語インデックス i' の組からなる単語アライメント $A = \{\langle i_1, i'_1\rangle, \ldots, \langle i_{|A|}, i'_{|A|}\rangle\}$ が与えられ、これに基づいてフレーズ対の集合 $BP = \{\langle f_{i_1}^{j_1}, e_{i'_1}^{j'_1}\rangle, \ldots, \langle f_{i_{|BP|}}^{j_{|BP|}}, e_{i'_{|BP|}}^{j'_{|BP|}}\rangle\}$ を抽出する。ここで、$f_{i_1}^{j_1}$ は i_1 単語目から j_1 単語目を含む f_1^J の部分単語列であり、$e_{i'_1}^{j'_1}$ は目的言語に対して同様に定義される。あるフレーズ対 $\langle f_i^j, e_{i'}^{j'}\rangle$ が抽出される条件は、両方のフレーズ中に、対応する単語が 1 つ以上含まれており、1 つのフレーズに含まれながらもう一方のフレーズに含まれない単語対応が存在しないことである。この条件に合致するフレーズは Och[9] の `phrase-extract` アルゴリズムで抽出できる。フレーズから同期ルールを作成するために、BP 中の各 $\langle f_i^j, e_{i'}^{j'}\rangle$ に対して、そのフレーズに含まれる部分フレーズを列挙し、非終端記号で置き換える [4]。

このルール抽出を複数の目的言語を扱える MSCFG へと拡張するために、まず原言語コーパス $\mathcal{F}$ と目的言語コーパス $\{\mathcal{E}_1, \ldots, \mathcal{E}_N\}$ を準備し、各 $\mathcal{F}$ と $\mathcal{E}_n$ の組に対して独立に単語アライメントを行う。この結果、原言語文 F に対して N 個の目的言語文 $\{E_1, \ldots, E_N\}$ と N 個のアライメント $\{A_1, \ldots, A_N\}$ が得られる。次に、N 個の言語対に対して、通常の 2 言語に対する `phrase-extract` アルゴリズムを用いて、N 個のフレーズ対集合 $\{BP_1, \ldots, BP_N\}$ を抽出する。

最後に、この N 個のフレーズ対集合を、すべての目的言語を同時に考慮した多言語フレーズ集合に変換する。具体的には、各原言語フレーズ f_i^j に着目し、各フレーズ集合 BP_n において f_i^j と対応する 0 個以上の目的言語フレーズ対集合を TP_n とする。すべての TP_n の直積集合を作成し、これに f_i^j を加えることで、多言語フレーズ集合を得る。つまり、f_i^j は T1 において 2 つのフレーズと対応され、T2 において 3 つのフレーズと対応されるなら、f_i^j に対して $2 * 3 = 6$ 個のフレーズ三つ組を抽出することとなる。このようにすべての目的言語を考慮した多言語フレーズを抽出してから、通常の SCFG と同じく、部分フレーズを (原言語とすべての目的言語において) 非終端記号に置き換え、MSCFG ルールを得る。

4.2 ルールのスコア付け

ルールを抽出してから、各ルールの品質を評価する素性関数を計算する。2 言語の記号列 γ と α_1 を有する通常の SCFG ルールでは、翻訳確率 $P(\gamma|\alpha_1)$ と $P(\alpha_1|\gamma)$ の対数、語彙化翻訳確率 $P_{lex}(\gamma|\alpha_1)$ と $P_{lex}(\alpha_1|\gamma)$ の対数、α_1 の終端記号数 (単語数)、フレーズペナルティなど、6 つの素性を計算するのが一般的である。

MSCFG では、さらに α_1 以外の目的言語記号列を考慮した素性を定義することができる。特に、本研究の実験では $N = 2$ の場合を扱うため、2 つ目の目的言語記号列 α_2 を含めた翻訳確率 $P(\gamma|\alpha_2)$ と $P(\alpha_2|\gamma)$ の対数、語彙化翻訳確率 $P_{lex}(\gamma|\alpha_2)$ と $P_{lex}(\alpha_2|\gamma)$ の対数、α_2 の終端記号数を素性として定義する。また、両方の目的言語記号列を考慮した翻訳確率 $P(\gamma|\alpha_1, \alpha_2)$ と $P(\alpha_1, \alpha_2|\gamma)$ の対数も用いる。これらを合わせて、2 つの目的言語を考慮した MSCFG では 13 個の素性が得られる。

4.3 ルールのフィルタリング

機械翻訳では、訳出に必要な計算量とメモリ領域を削減するために、事前に翻訳モデルから一部のルールをフィルタリングするのが一般的である。例えば、各 γ に対して $P(\alpha_1|\gamma)$ の順に α_1 の候補を並べ、確率最大の L 個のルールのみを訳出に利用する手法が広く使われる。

本研究では、このフィルタリング法を 2 つの目的言語を考慮した MSCFG に適応する 2 つの手法を提案し、実験で比較する。1 つ目の手法を **T1+T2 フィルタリング法**と定義し、両方の目的言語を考慮した翻訳確率 $P(\alpha_1, \alpha_2|\gamma)$ に基づいて、確率最大の L 個のルールを選択する。このフィルタリング法は通常の SCFG における手法の単純な拡張であるが、各 α_1 に対して多くの α_2 が存在する場合、フィルタリング後に残る α_1 の候補数を減らしてしまう可能性もある。この問題に対応するために、まず $P(\alpha_1|\gamma)$ を用いて確率最大の α_1 の候補を L 個選択してから、各 α_1 に対して確率 $P(\alpha_1, \alpha_2|\gamma)$ が最大の α_2 を選択する **T1 フィルタリング法**も提案する。これにより必ず α_1 の多様性は保証されるが、その一方 α_1 に対して α_2 が 1 つしか残らないという強い制約も課される。

5 複数の言語モデルを考慮した探索

LM は目的言語文の確率 $P(E)$ を計算し、通常の翻訳においても提案法においても必要不可欠な情報源である。しかし、LM 確率は 4.2 節で紹介されたような各ルールに対して独立に計算する局所素性とは異なり、ルールの組み合わせに依存する。n-gram の LM 確率

$$P_{LM}(E) = \prod_{i=1}^{|E|+1} p(e_i|e_{i-n+1}, \ldots, e_{i-2}, e_{i-1}) \qquad (4)$$

の場合、各単語 e_i の確率は直前の $n-1$ 単語に依存する。

LM を使用しない場合の CKY+アルゴリズム [3] は、原言語スパン f_i^j の最もスコアの高い訳出候補を記憶す

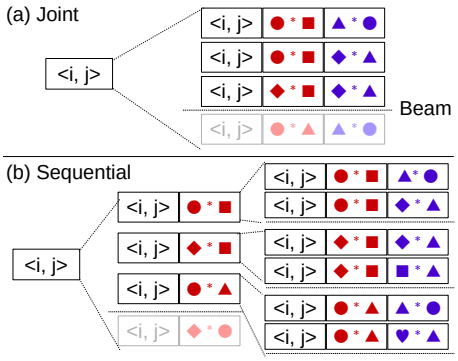

図 2: (a) 同時探索と (b) 逐次探索の状態の例

る動的計画法に基づく。LM 確率を考慮する際、f_i^j に対して、更に訳出候補の両側の $n-1$ 個の単語も状態として区別する必要がある。この状態の拡張で探索空間が膨らみ、現実的な時間での全探索が不可能となるため、各スパンにおいて考慮する組み合わせの数に対して制限 (pop limit) を設け、近似的な探索を行う [4]。複数の目的言語文に対する LM を考慮する場合、各言語の両端の $n-1$ 単語を考慮する必要が生じ、さらに探索空間が広がる。この探索を行う方法を 2 つ提案する。

まず、**同時探索**では、T1 と T2 の両端の単語を同時に考慮した探索を行う。図 2(a) に示すように、両言語の単語を考慮した LM 状態を定義し、これを通常の手続きで展開する。既存の探索法の拡張で比較的シンプルである反面、それぞれの言語で考慮する候補の数が、同じビームを用いた単言語の探索に比べて減少する問題もある。例えば、図では 3 つの仮説が展開されたにも関わらず、T1 の単語列に重複があるため、T1 の単語列の異なり数は 2 にとどまる。この多様性の減少は翻訳精度に悪影響を及ぼす恐れがある。

この問題の解決策の 1 つとして、**逐次探索**も提案する。逐次探索では、図 2(b) に示すように、まず T1 を考慮した探索を行ってから、T2 を考慮する探索を行う。これにより、仮説の T1 における多様性は保証されるが、T2 の LM 確率が探索に大きな影響を与える場合、T1 のみを用いた探索の段階で失敗する要因にもなり得る。

6 実験的評価

6.1 実験設定

実験の題材として、国連文章から構成される MultiUN コーパスを用いる [5]。このコーパスを題材に選んだのは、アラビア語 (ar)、英語 (en)、スペイン語 (es)、フランス語 (fr)、ロシア語 (ru)、中国語 (zh) という言語的特徴の異なる大量な対訳データを含むためである。元文章が英語の場合が多いため、実験でも英語を原言語とし、それ以外の言語を第 1 と第 2 の目的言語とする。まず、文のアライメントを取り、2 回以上コーパスに現れる原言語文を 1 回のみに減らして、約 350 万文からなる、6 カ国語対訳コーパスを作成する。この中から、1,500 文ずつパラメータ調整とテストデータとして取り除き、残りを学習データとする。また、1 節で述べたような、T1 の言語資源が少ない状況を想定して、翻訳モデルと T1 の LM の学習データを 10 万文に限定し、T2 の LM 学習をすべての約 350 万文で行った。

T2 \ T1	ar	es	fr	ru	zh
-	24.77	42.07	37.22	26.18	21.14
ar	-	40.88	37.44	25.96	21.13
es	†**25.22**	-	‡**38.85**	**26.47**	**21.37**
fr	25.02	‡**42.84**	-	26.19	21.25
ru	24.52	41.94	†37.80	-	21.15
zh	24.19	41.95	37.21	25.67	-

表 1: 通常の SCFG（1 行目）と MSCFG が T2 を考慮した場合（2〜6 行目）の翻訳結果。太字は各 T1 における最も高い精度であり、短剣符は SCFG ベースラインとの有意差を示す (†: $p < 0.05$, ‡: $p < 0.01$)

SCFG と MSCFG の翻訳システムを、Travatar[8] を用いて作成した。T2 を導入することで T1 の翻訳精度向上を実現する本研究の主な目的を考慮して、モデルの最適化と評価に T1 の BLEU スコアを利用した。断りのない限り、探索にはビーム幅 2,000 の同時探索を用い、T1 フィルタリング法で確率上位の 10 個のルールに制限した。なお、事件結果の議論の際、目的言語を「T1/T2」(例：fr/es) の形式で記述する。

6.2 複数の目的言語を考慮する効果

まず本節で、複数の目的言語を考慮することで、T1 の翻訳精度向上を実現できるかを検証する。5 つの言語を T1 とし、残りの言語をそれぞれ T2 にした結果と、ベースラインである通常の SCFG の結果を表 1 に示す。この図から、すべての T1 において、T2 を考慮した方が高い性能が見られ、特に英語に言語的構造の近いスペイン語は（T1 がスペイン語の場合を除いて）すべての言語対で最も有効な T2 であった。このことから、複数の目的言語を考慮した提案法は確かに性能の向上に貢献できることが分かる。

上がり幅を見てみると、fr/es が最も高い 1.63 点を記録し、es/fr の 0.77 点が続く。その一方で、中国語やロシア語が含まれる言語対では、有意差が見られず、若干の性能悪化が見られることもあった。このことから、提案法は特に言語的特徴の類似している言語を対象とした場合に有効であると言える。

6.3 T1 の LM 性能、フィルタリング、探索の影響

T1 の LM 性能：今までの実験で、T1 の LM 学習データを 10 万文に限定することで、T2 の LM より弱いモデルを構築した。T2 の LM は、T1 の弱い LM を補う役割を果たしているため、性能の上がり幅も T1 の LM の相対的な弱さに依存すると仮定できる。この影響を調べるために、T1 の学習データを 0 文から全データの 350 万文へと変動させ、各データ数での T2 の追加による向上幅を調べた。図 3 に、特に影響が顕著だった fr/es の実験を示す。この結果、0 文の場合に T2 の導入により 3.79 点という大幅な性能向上が見られ、T1 の学習に用いる文数が増えるほど上がり幅が狭まり、40 万文程度を T1 の学習に用いた場合はほぼ差がなくなっている。このことから、T1 の LM に比べて T2 で性能の良い LM が作

T2	T1=fr T1+T2	T1=fr T1	T1=zh T1+T2	T1=zh T1
ar	36.12	**37.44**	20.57	**21.13**
es	**38.87**	38.85	20.47	**21.37**
fr	-		19.90	**21.25**
ru	37.02	**37.80**	20.04	**21.15**
zh	37.06	**37.21**	-	

表 2: フィルタリング法による翻訳精度の差

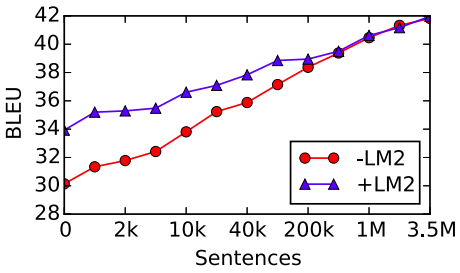

図 3: 様々なデータサイズで学習した T1 LM に対する T2 の LM なし (-LM2) とあり (+LM2) の翻訳精度

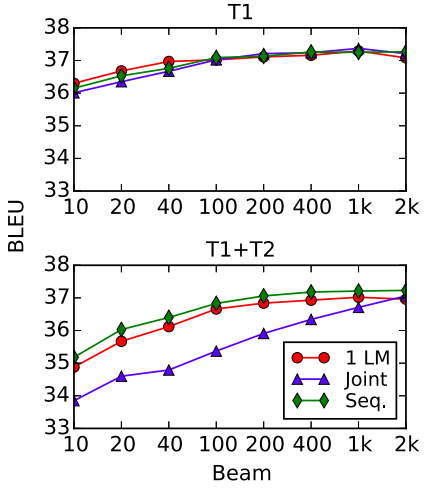

図 4: T2 の LM なし (1 LM)、同時探索 (Joint)、逐次探索 (Seq.) の翻訳精度

成可能な場合に、提案法は特に有効であると言える。

フィルタリング法：4.3 節で述べたフィルタリング法の影響を調べるために、各フィルタリング法を用いた翻訳実験を行った。T1 をフランス語と中国語とし、確率最大の 10 個のルールを用いた場合の実験結果を図 2 に示す。この実験結果から、T1 フィルタリング法はほとんどの場合において、高い翻訳精度を実現していることが分かり、4.3 節で仮定された通り、T1 の多様性を維持することが T1 の翻訳精度向上につながることが分かる。その一方、唯一の例外として fr/es の結果が挙げられ、T1 と T2 が十分に類似している場合はフィルタリング法の工夫はさほど必要ないことも分かる。

探索法：最後に、同時探索と逐次探索の影響を調べた。fr/zh の翻訳方向においてフィルタリング法を T1 法もしくは T1+T2 法とし、ビーム幅を様々な値と変動させた。逐次探索の場合の T2 のビーム幅は大きな影響がなかったため、10 と固定した。また、T2 の LM を用いないシステムの結果も参考として示す。

図 4 の実験結果から分かる通り、T1 フィルタリング法を用いる場合、両方の探索法に大きな差が見られないが、T1+T2 フィルタリング法を用いる場合、逐次探索法において大幅な性能向上が見られる。このことから、フィルタリングにより T1 の多様性が保証されている場合は逐次探索の工夫は必要ないが、単純な T1+T2 フィルタリングなら逐次探索は T1 の多様性を保持し、性能の向上につながることが分かる。なお、紙面の都合上詳細な結果を示さないが、fr/es の方向で同様の実験を行った際、いずれのフィルタリング法でも探索法の精度に大差がなかったことから、T1 と T2 が類似している場合、この工夫の必要性が減少することも言える。

7 おわりに

本研究では、複数の目的言語を考慮した新たな機械翻訳枠組みを提案し、第 2 の目的言語の自然性を評価することで第 1 の目的言語の翻訳精度向上が実現可能であることを示した。今後の課題として、全言語を含んだ対訳データが入手可能でない場合の学習法、複数の言語を考慮した単語アライメント [6] との組み合わせ、両目的言語の翻訳精度を最適化する枠組みなどが考えられる。

謝辞：有益な助言を下さった渡辺太郎氏と三浦明波氏に感謝します。本研究の一部は、Microsoft CORE プロジェクトおよび JSPS 科研費 25730136 の助成を受け実施したものである。

参考文献

[1] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In *Proc. EMNLP*, pages 858–867, 2007.

[2] M. Cettolo, C. Girardi, and M. Federico. WIT3: web inventory of transcribed and translated talks. In *Proc. EAMT*, pages 261–268, 2012.

[3] J.-C. Chappelier, M. Rajman, et al. A generalized CYK algorithm for parsing stochastic CFG. In *TAPD*, pages 133–137, 1998.

[4] D. Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

[5] A. Eisele and Y. Chen. MultiUN: A multilingual corpus from United Nation documents. In *Proc. LREC*, 2010.

[6] A. Lardilleux and Y. Lepage. Sampling-based multilingual alignment. In *Proc. RANLP*, pages 214–218, 2009.

[7] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proc. EMNLP*, pages 880–889, 2009.

[8] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proc. ACL Demo Track*, pages 91–96, 2013.

[9] F. J. Och. *Statistical machine translation: from single-word models to alignment templates*. PhD thesis, RWTH Aachen, 2002.

[10] F. J. Och and H. Ney. Statistical multi-source translation. In *Proc. MT Summit*, pages 253–258, 2001.

[11] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.