原言語側対訳データを要さない 統計的機械翻訳法への中間言語モデルの導入

楠本 高康　　秋葉 友良
豊橋技術科学大学 情報・知能工学専攻
{kusumoto, akiba}@.cl.ics.tut.ac.jp

1　はじめに

情報化社会の進展とともに、外国語の情報に触れる機会が増加する中で、機械翻訳技術の向上が望まれている。統計的機械翻訳 (SMT) は、対訳コーパスと呼ばれる大量の対訳文から翻訳規則を機械学習する手法であり、言語間の専門的な知識を必要とせずに構築することが出来る。しかし、対訳コーパスの作成には人的・時間的なコストがかかるため、言語的な資源に乏しく、十分な量の対訳コーパスを利用できないような言語間で SMT を構成することは難しい。

対訳コーパスが利用できない原言語-目的言語間を翻訳するために、原言語と目的言語両方との間に対訳コーパスをもつ中間言語を利用する手法が提案されている [6]。しかし、この手法でも原言語-中間言語間の対訳コーパスを必要とするため、どの言語間にも十分な量の対訳コーパスが利用できないほど言語資源が限られている原言語を翻訳することは出来ない。

対訳コーパスが利用できない原言語を翻訳するために、Manh らは原言語-中間言語の対訳コーパスの代わりに単語辞書を利用する手法を考案し、ベトナム語-日本語翻訳の性能を調査した [5]。

本研究では、中間言語の言語モデルを利用することで、Manh らの手法の性能の改善を試みた。また、提案手法と、原言語-目的言語間で直接対訳コーパスを用いて作成した手法、および二種類の対訳コーパス（原言語-中間言語, 中間言語-目的言語）を用いて作成した手法の性能を比較した。

2　提案手法

単語辞書を用いた翻訳には、訳語の選出や並び替えといった課題がある。Manh らはこれらの課題に取り組むために、単語辞書の全ての訳語の組み合わせを表現するワードラティスを作成し、ラティスデコーダを用いて翻訳確率が最も高くなる組み合わせを選んだ。また、対訳コーパスから学習された中間言語のフレーズを事例として利用し、語の並び替えを行った。筆者らは更に同様の手法を用いて中間言語に特有な語を補間した。

本研究では、さらに中間言語の言語モデルを利用して、ラティスの中でより中間言語らしい文に高い重みを付けることで、より中間言語文らしい文に高い翻訳確率を与えた。

2.1　ラティスの作成

単語辞書から適切な訳語を選択するために、複数の文の候補を表現することが可能なワードラティスと、ワードラティスを入力として用いることが出来る統計的機械翻訳システム [1] を利用した。頂点 v から出発して頂点 w に接続する、ラベル l を持つ辺を (v, w, l) として表すと、単語辞書の訳語の組み合わせ全てを表現するワードラティスは以下ようにして作成することが出来る。

入力: 分割された原言語の入力文 $f = f_1, f_2, \cdots, f_n$.

出力: 有向グラフ $\Sigma = (V, E)$.

1. $V = \{v_0\}, E = \{\}$ として初期化する。
2. For $i = 1$ to n:
   (a) 頂点 v_i を作り、V を $V \leftarrow V \cup \{v_i\}$ で更新する。
   (b) 単語辞書から単語 f_i の訳の集合 S を得る。このとき訳が見つからなければ f_i 自身を訳として用いる。
   (c) $s \in S$ について:
   s が単語であれば、辺 (v_{i-1}, v_i, s) を作成して E に追加する。
   s が k 語のフレーズ $w_1, w_2, \cdots w_k$ であれば、k　1 個の新たな頂点 $v'_1 \cdots v'_{k-1}$ と k 個の辺 (v_{i-1}, v'_1, w_1), (v'_1, v'_2, w_2), $\cdots$, (v'_{k-1}, v_i, w_k) を作成し、それぞれ V と E に追加する。

上述のステップで作成されたワードラティスの例を図 1 に示す。

ラティスデコーダにより、このワードラティスから目的言語の翻訳確率が最も高くなる単語の組み合わせが選択される。

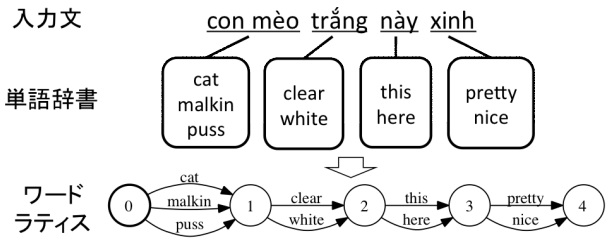

図 1: ワードラティスの作成

2.2 ラティスの拡張

前節で作成したラティスから生成される文の語順は原言語と同じであるので、単語を適切に並び替える必要がある。同様に、中間言語の冠詞などに対応する単語が原言語文に含まれていない場合、適切な語を補った。

単語列を並び替えたり、単語を補完すべき箇所を判断するために、中間言語-目的言語の対訳コーパスから学習した、中間言語のフレーズを利用した。ワードラティス中のあらゆる n-gram$(n = 3, 4, 5)$ の部分文字列 S を探索し、S を並び替えたり、S に語の補間を行った文字列が SMT の学習データに含まれていた場合、その事例を新たにワードラティスに追加することで、語の並び替え及び補間をする。こうして追加された新たなパスは機械翻訳システムの構成要素の一部であるため、翻訳の際に選ばれやすくなる。

2.2.1 単語の列び変え

ワードラティスの部分文字列を探索し、フレーズテーブルにそれを並び替えたものが含まれていた場合、そのフレーズをワードラティスに追加することで、単語の並び替えを行った。

以下の手順に従って、フレーズをワードラティスに追加した。

入力: ワードラティス $\Sigma = (V, E)$, フレーズテーブル T

出力: 拡張されたワードラティス $\Sigma = (V', E')$

1. $V' = V, E' = E$ として初期化する。
2. $v_i \in V$ について:
   (a) 始点 v_i の n-gram パス $p_i^{i+n} = ((v_i, v_{i+1}, w_1), (v_{i+1}, v_{i+2}, w_2), \cdots, (v_{i+n-1}, v_{i+n}, w_n))$ の集合 P を集める。
   (b) $p \in P$ の単語列 “$w_1 w_2 \cdots w_n$” を並び替えた単語列が T に含まれるフレーズ “$x_1, x_2, \cdots, x_n$” と等しければ、
      i. 新たな頂点 $V'' = \{v'_1, v'_2 \cdots, v'_{n-1}\}$ 及び辺 $E'' = \{(v_i, v'_1, x_1), (v'_1, v'_2, x_2), \cdots, (v'_{n-1}, v_{i+n}, x_n)\}$ を作成する。
      ii. $V' \leftarrow V' \cup V''$
      $E' \leftarrow E' \cup E''$ で更新する。

フレーズの追加による単語の並び替えの例を図 2 に示す。図では、 “cat white this” という単語列を並び替えたフレーズ “this white cat” を追加することで単語の並び替えをおこなっている。

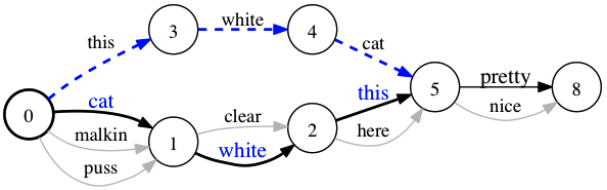

図 2: 単語並び変えの例

2.2.2 単語の補完

2.2.1 節と同様の手法を用いて、以下の手順に従って中間言語に固有な単語の補間を行った。

入力: ワードラティス $\Sigma = (V, E)$、フレーズテーブル T 及び補完する単語のリスト L.

出力: 拡張されたワードラティス $\Sigma = (V', E')$.

1. $V' = V, E' = E$ に初期化する。
2. $v_i \in V$ について:
   (a) 始点 v_i の n-gram パス $p_i^{i+n} = ((v_i, v_{i+1}, w_1), (v_{i+1}, v_{i+2}, w_2), \cdots, (v_{i+n-1}, v_{i+n}, w_n))$ の集合 P を集める。
   (b) $p \in P$ の単語列に、L に含まれる単語を 1 語加えた文字列 “$w_1\ w_2 \cdots w_j\ y\ w_{j+1} \cdots w_n$” $(y \in L\ , 1 \leq j \leq m)$ が、T に含まれるフレーズ $x_1, x_2, \cdots x_{n+1}$ と等しければ：
      i. 新たな頂点 $V'' = \{v'_1, v'_2, \cdots v'_n\}$ 及び辺 $E'' = \{(v_i, v'_1, x_1), (v'_1, v'_2, x_2), \cdots, (v'_{n-1}, v'_n, x_n), (v'_n, v_{i+n}, x_{n+1})\}$ を作成する
      ii. $V' \leftarrow V' \cup V''$
      $E' \leftarrow E' \cup E''$ で更新する。

フレーズの追加による単語の補間の例を図 3 に示す。図では、 “cat pretty” という単語列に語を補ったフレーズ “cat is pretty” を追加することで単語の並び替えをおこなっている。

2.3 ラティスの重み付け

ラティスデコーダがパスを探索する際、中間言語文らしい文を入力として用いた時にボーナスを与えるた

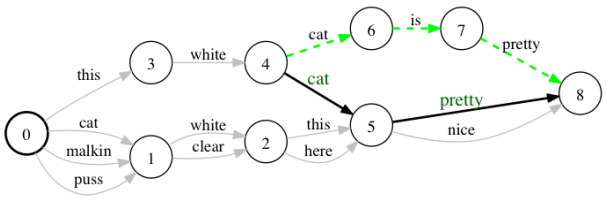

図 3: 単語補間の例

めに、中間言語の言語モデルを用いてラティスの各辺に重みを加えた。

$p(e)$, $n(e)$, $l(e)$ は、それぞれ辺 $e = (v, w, l)$ の各要素とし、辺 e の重みを $w(e)$ とする。$E_{in}(v)$ は頂点 v へ入る辺の集合 $\{e \in E | n(v) = n\}$ であり、$V_{in}(v)$ は頂点 v へ入る頂点の集合 $\{p(e) \in E_{in}(v)\}$ である。頂点 v からでる頂点と辺の集合 $V_{out}(m)$ と $E_{out}(m)$ をそれぞれ同様に定義し、以下のようにしてラティスの重み付けを行った。

入力: ワードラティス $\Sigma = (V, E)$

出力: 重み付けされたワードラティス $\Sigma = (V', E')$.

1. $V' = V, E' = E$ として初期化する。
2. $e \in E$ について、3-gram 言語モデル確率 $\max_{e_{-1}, e_{-2}} P(l(e)|l(e_{-1}), l(e_{-2}))$ を与える。
3. 頂点 $n \in V$ ごとに、前向き確率 $\alpha(n)$ と後向き確率 (n) を求める。

$$\alpha(n) = \left(\sum_{n_{-1} \in V_{in}(n)} \alpha(n_{-1}) \right) \left(\sum_{e \in E_{out}(n)} w(e) \right)$$

$$(n) = \left(\sum_{n_{+1} \in V_{out}(n)} (n_{+1}) \right) \left(\sum_{e \in E_{in}(n)} w(e) \right)$$

4. α と を用いて、各辺に言語モデルの事後確率 e'_w を与える。

$$e'_w = \frac{\alpha(u) \cdot w(e) \cdot (v)}{(0)}$$

ラティスから選ばれたパスを $F = e_0, e_1, \cdots, e_N$ として、言語モデルを利用した各辺の重みの素性 f_l 及び、標準的な SMT のワードペナルティに類似した素性 f_p を以下のように定義する。

$$f_l(E, F) = \sum_{i=0}^{N} \log w(e_i) \quad (1)$$

$$f_p(E, F) = |F| \quad (2)$$

f_l を翻訳モデルに加えることで、より中間言語文らしい文に高い翻訳確率を与え、f_p を利用して、ラティスから適切な長さの中間言語文を選択した。

標準的な SMT システムの素性 (フレーズ翻訳確率、語歪み確率など) に f_l と f_p を加えて、対数線形モデルで翻訳確率を計算した。翻訳確率が最大になる目的言語文 $\hat{E}$ は、以下の式で求めることが出来る。

$$\hat{E} = \arg \max_{E} \max_{F \in O} \sum_{n=1}^{N} \lambda_n f_n(E, F) \quad (3)$$

ここで、E は目的言語文、O は入力のラティス、F は O から選ばれた中間言語文（あるいはパス）である。f_n, λ_n はそれぞれ翻訳確率を計算するための素性とその重みである。

3 実験

原言語-目的言語間の対訳コーパスを直接用いて作成したシステムや、原言語-中間言語，中間言語-目的言語間の二種類の対訳コーパスを利用したシステムと比較するために、フランス語を資源の制限された言語であるとみなして、フランス語-スペイン語間で翻訳を行った。どちらの実験でも、見出し語化した英語を中間言語として用いて、ラティスが生成する文の数が 10^5 個以内になるように枝狩りをした。ラティスデコーダとして Moses [3] を用いた。

3.1 実験データ

Europarl コーパス [4] の Release v7 で配布された、英語-フランス語コーパスと、英語-スペイン語コーパスから、内山らの手法 [6] に従ってフランス語-英語-スペイン語の 3 言語パラレルコーパスを作成した。

パラレルコーパスのうち 60 万文を訓練セット、1000 文を開発セット、100 文をテストセットとして用いた。テストセットには、原言語（フランス語）側の文長が 20 文字以内のものを選んだ。

訓練セットを用いて翻訳モデルを訓練し、開発セット 1000 文の BLEU スコアを最大化するように f_l, f_p を除く各種の重みを MERT で最適化した。f_l と f_p の重みは、フランス語文の長さが 20 文字以内の開発セットの 300 文を用いてそれぞれ調整した。スペイン語言語モデルとして、訓練データのスペイン語側を用いて irstlm [2] で訓練した 5-gram 言語モデルを用いた。

単語辞書として、フランス語-英語対訳コーパスから学習したフレーズテーブルのエントリを単語辞書として利用した。ソースフレーズとターゲットフレーズのに含まれる単語数の差が 2 以内のものを辞書のエントリとして用いた。翻訳語が 5 つ以上あった場合、フレーズ翻訳確率が高いものから上位 5 つを翻訳語として用いた。

3.2 実験結果

ベースライン 標準的な原言語-目的語間の対訳コーパスを用いて訓練したSMTシステム (**DIRECT**) と、原言語と目的言語両方との対訳コーパスをもつ中間言語を利用するシステム (**SENT15**) をベースラインとして用いた。**SENT15** は原言語-中間言語の機械翻訳結果の15-bestのうち目的言語への翻訳確率が最大になる文を選んで最終出力とするシステムである [6]。

比較手法 単語辞書を用いて原言語を中間言語に翻訳する際に、単語辞書の先頭にある単語を選ぶシステム（**TOP_W**）と、ラティスから言語モデルが最大になるパスを選択するシステム (**TOP_LM**) の性能を比較した。また、語を並び替えるフレーズ、語を補完するフレーズを追加した場合としなかった場合で翻訳結果を比較した。

表1に各システムの翻訳結果のBLEUスコアを示す。f_l と f_p に調整した重みを与えた場合と、重み0を与えた場合でそれぞれBLEUスコアを求めた。

表 1: フランス語-スペイン語翻訳 BLEU スコア

提案手法		
フレーズ拡張	英語言語モデル	
	なし	あり
拡張なし	22.02	20.18
並び変え	19.31	19.33
補間	22.07	19.76
両方	20.05	19.53

ベースライン	
DIRECT	37.00
SENT15	30.70
TOP_W	2.43
TOP_LM	20.78

表1より、単語辞書の先頭を中間言語文として選ぶシステムと比較して、ラティスから言語モデルが最大になるパスを用いたモデルのBLEUスコアが大きく改善されていることがわかった。ここから、中間言語の言語モデルは、ラティスから中間言語を選ぶ際に有用であると考えられる。一方で、既存のシステムに中間言語らしさを重みとして与えた場合では、性能の改善が見られなかった。

この原因を調査するため、f_l と f_p を重みとして用いた場合と用いなかった場合で、最終的な翻訳結果に用いられた中間言語文がどのように変化したかを比較した。その結果、f_l と f_p を重みとして用いた場合は、中間言語文の文長が平均20単語以上長くなっていることがわかった。この原因は、開発セットとして用いた文が少なく、重みの調整が不完全であったからだと考えている。MERTを用いて f_l と f_p の重みをより適切に調整することで、システムの性能を改善できると考えている。

4 おわりに

利用可能な言語資源が限られている原言語を翻訳するため、原言語を含む対訳コーパスの代わりに、言語資源が豊富な中間言語の単語辞書を利用した。実験結果から、直接対訳コーパスを持つシステムと比較して、最大73%の性能を発揮することがわかった。今後の予定として、本論文で提案した素性 f_l と f_p への重みを適切に調節して実験することを考えている。

参考文献

[1] Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pp. 1012–1020, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[2] Marcello Federico and Mauro Cettolo. Efficient handling of n-gram language models for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 88–95, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] Hieu Hoang, Alexandra Birch, Chris Callisonburch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ond?ej Bojar. Moses: Open source toolkit for statistical machine translation. pp. 177–180, 2007.

[4] P. Koehn. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, Vol. 5, , 2005.

[5] 秋葉友良 Nguyen Manh Hung. Word lattice decoding を利用した対訳コーパスのない言語からの統計的機械翻訳. 言語処理学会 第16回年次大会発表論文集, pp. 1006–1009, 2009.

[6] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 484–491, Rochester, New York, April 2007. Association for Computational Linguistics.