パラフレーズラティスを用いた統計的機械翻訳

大西 貴士　　内山 将夫　　隅田 英一郎
情報通信研究機構　MASTAR プロジェクト　言語翻訳グループ
{takashi.onishi,mutiyama,eiichiro.sumita}@nict.go.jp

1　はじめに

統計的機械翻訳 (SMT) におけるラティスデコーディングは，音声翻訳や中国語翻訳等で高精度な翻訳が得られることが知られている [1] [2]．例えば，音声翻訳では音声認識結果を入力として翻訳が行われるが，1-best の認識結果だけでなくその他の認識候補もまとめてラティスを作り，それを入力としてラティスデコーディングを適用することで，音声認識による曖昧性を考慮した翻訳が行えるようになる．そのため，翻訳精度は 1-best を入力としたものよりもラティスを入力としたもののほうが高くなる．

本論文では，原文のパラフレーズを扱う上でもラティスデコーディングが有効であることを示す．パラフレーズとは，言い換え表現のことで，表層的には異なるが意味的には同等である表現のことである．自然言語ではこのようなパラフレーズは数多く存在する．そのため，たまたま原文での表現がトレーニングコーパスに含まれていなかったり，出現頻度が少なかったりすることによって翻訳が失敗するという課題がある．そこで，翻訳の原文として与えられたものが数多くあるパラフレーズの中の 1 つ (1-best) であると考え，原文を直接翻訳するのではなく，原文をパラフレーズすることで自然言語の曖昧性をラティスの形で表現し，そのラティスに対してラティスデコーディングを適用する手法を提案する．

パラフレーズを利用することで，元の原文中にトレーニングコーパスにない表現があってもパラフレーズした表現がトレーニングコーパスにあれば正しく翻訳することができる．また，ラティスデコーディングを利用することで，デコーディング時の素性に原言語側の言語モデル情報を組み入れることができ文脈に応じて適切なパラフレーズを選び翻訳することができる．

以下では，2 章で関連研究を紹介し，3 章で本論文で提案するパラフレーズラティスを用いた翻訳手法について説明する．4 章で IWSLT2007 のデータセットを用いて行った実験について述べ，5 章でまとめと今後の展開について述べる．

2　関連研究

ラティスを用いた SMT として，音声翻訳において，音声認識結果をラティスの一種である Confusion Network で表し，それを入力として翻訳を行うもの [1] や，中国語翻訳において，中国語の形態素解析の曖昧性をラティスで表し，それにラティスデコーディングを適用するもの [2] 等がある．これらは，翻訳の前段階で行う処理に関する曖昧性をラティスデコーディングで解決する手法であるが，パラフレーズに対してラティスデコーディングを適用したものはなかった．

一方，パラフレーズを用いた SMT としては，パラフレーズによって未知語（未知フレーズ）の翻訳を獲得し，それをフレーズテーブルに追加することで未知語（未知フレーズ）の翻訳を行うもの [3] や，パラフレーズによってトレーニングコーパスを展開するもの [4] 等がある．しかし，パラフレーズを用いた SMT で，原文をパラフレーズし，それとラティスデコーディングと組み合わせたものはなかった．

3　パラフレーズラティスを用いた SMT

提案手法での翻訳の流れは図 1 のようになる．あらかじめ，トレーニング用とは別のパラレルコーパスからパラフレーズを自動的に獲得しておく．原文が与えられると，獲得したパラフレーズを用いて原文をパラフレーズし，ラティスの形式に変換する．このラティスのことをパラフレーズラティスと呼ぶ．最後に，こ

のパラフレーズラティスに対してラティスデコーディングを適用し翻訳文を得る．

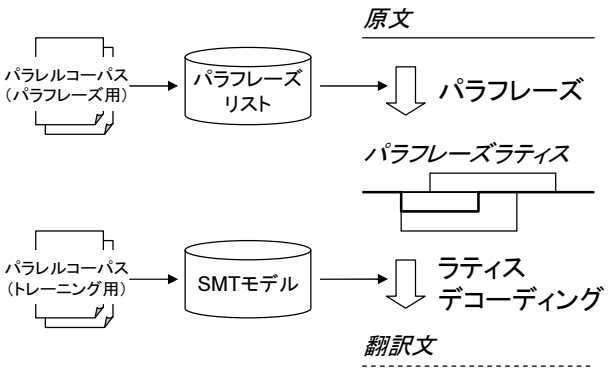

図 1: 翻訳の流れ

3.1 パラフレーズの獲得

Bannard ら [5] と同様の手法を用いてパラレルコーパスからパラフレーズを自動的に獲得する．これは，パラレルコーパスのアライメントをとり，ある言語の 2 つのフレーズ e_1，e_2 が双方とも別の言語のフレーズ c とアライメントされているなら，2 つのフレーズ e_1，e_2 がパラフレーズ候補であるとする手法で，手順は以下のようになる．

1. フレーズテーブルの作成
   パラレルコーパスから通常のフレーズベース SMT と同様の手順でフレーズテーブルを作成する．

2. sigtest-filter によるフィルタリング
   1 で得られたフレーズテーブルには信頼度の低いフレーズペアも含まれているため，sigtest-filter [6] を用いて信頼度の高いフレーズペアだけを残す．

3. パラフレーズ確率の算出
   e_1 のパラフレーズ候補として，e_2 がある場合，以下のようなパラフレーズ確率 $p(e_2|e_1)$ を算出する．

$$p(e_2|e_1) = \sum_c P(c|e_1)P(e_2|c)$$

   ここで，$P(\cdot|\cdot)$ はフレーズ翻訳確率である．

4. パラフレーズの獲得
   $p(e_2|e_1) > p(e_1|e_1)$ となるフレーズ e_2 を e_1 のパラフレーズとして抽出する．

3.2 パラフレーズラティスの作成

原文が与えられると，前節で獲得したパラフレーズリストを用いて原文をパラフレーズし，パラフレーズラティスを作成する．パラフレーズラティスは，原文をパラフレーズし，それをラティスとして表現したものである．図 2 にパラフレーズラティスの例を示す．この例では，“give me some anodyne , please .” という原文に，“anodyne”=“pain killer” と “anodyne”=“sedative” の 2 つのパラフレーズが適用されるによって，

- “give me some anodyne , please .”
- “give me some pain killer , please .”
- “give me some sedative , please .”

の 3 つの文を表すパラフレーズラティスとなっている．

パラフレーズラティスの各ノードには，単語，次ノードまでの距離の他に，ラティスデコーディングで用いられる素性を持つ．本研究では，ラティスデコーディングの素性として，以下の 4 種類の素性の中から，(p)，(p, l)，(p, L)，(p, l, d) の 4 通りの組み合わせを用いた．

- パラフレーズ確率 (p)
  パラフレーズ獲得時のパラフレーズ確率．
  $p(e_2|e_1) = \sum_c P(c|e_1)P(e_2|c)$

- 言語モデルスコア (l)
  パラフレーズした原文と元の原文の言語モデル確率の比．
  $lm(after)/lm(before)$

- 正規化言語モデルスコア (L)
  文長（単語数）で正規化した言語モデル確率の比．
  $LM(after)/LM(before)$
  $LM(sent) = lm(sent)^{1/length(sent)}$

- パラフレーズサイズ (d)
  パラフレーズ前後での文長（単語数）の差．
  $exp(length(after) - length(before))$

言語モデルに関する素性 (l, L) は，同じパラフレーズでも原文の文脈によって値が異なる．そのため，これらの素性を加えることによって文脈に適合しないパラフレーズにペナルティを与え，正しいパラフレーズを優先して翻訳することができる．また，素性 L，d は，パラフレーズによって文長が短くなる場合に言語モデ

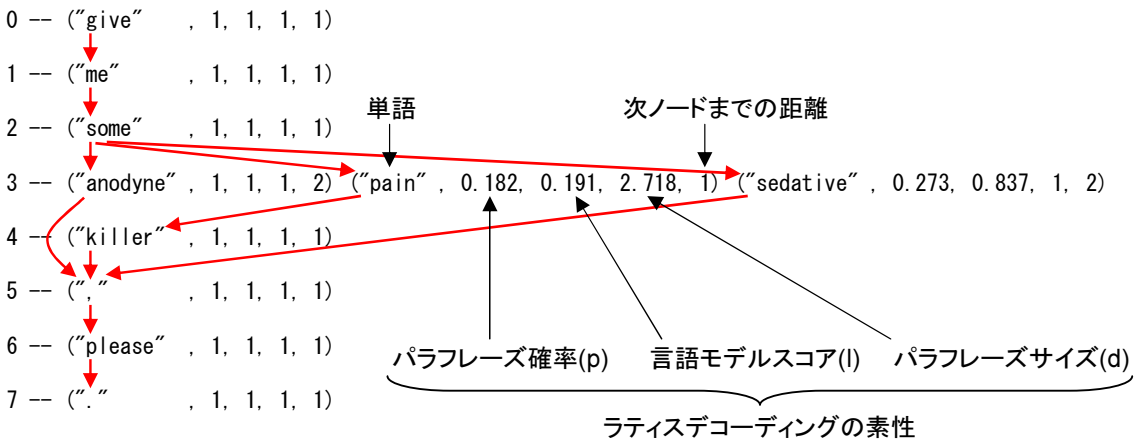

図 2: パラフレーズラティス

ルの値が有利になることを補正するために加えた素性である．

これらの素性は，パラフレーズによるノードの場合だけに付与し，その他のノードの場合は 1 とする．

3.3 ラティスデコーディング

ラティスデコーディングは，通常の SMT の素性とラティスの各ノードに付与された素性を用いてラティス中の最適なパスとそのときの翻訳結果を探索する．

本研究では，ラティスデコーディングを行うデコーダとして Moses [7] を用いた．Moses はオープンソースの SMT システムであり，ラティスを入力とすることができる．各素性の重みは MERT によってチューニングされる．

4 実験

提案手法の有効性を確認するために，英日，英中翻訳の実験を IWSLT2007 のデータセットを用いて行った．このデータセットは旅行会話に関するパラレルコーパスで，英日，英中ともトレーニング用として約 4 万文，訓練用，評価用として dev1〜dev3 セットが各々約 500 文となっている．dev1 は，パラメータチューニング用，dev2 は，後述する提案手法の条件設定の選択用，dev3 は，評価用として用いた．

パラフレーズの獲得は，英日翻訳の場合は英中のパラレルコーパスを用いて行い，約 5.3 万ペアのパラフレーズリストが得られた．同様に，英中翻訳の場合は英日のパラレルコーパスから約 4.7 万ペアが得られた．

4.1 ベースライン手法

ベースラインとして Moses と，Callison-Burch ら [3] の手法 (CCB) を用いた．Moses は，パラフレーズを行わず通常のフレーズベース SMT を行った．CCB では，フレーズテーブルを提案手法と同じパラフレーズリストを用いて展開し，それによって新たなフレーズ翻訳が得られると，それをフレーズテーブルに追加して翻訳を行った．このとき，フレーズテーブルにパラフレーズ用の素性としてパラフレーズ確率 (p) を追加し，MERT でパラメータチューニングを行った．

4.2 提案手法

提案手法では，パラフレーズラティス作成時の制限や，ラティスデコーディング時に用いる素性によって条件を変えて実験を行い，dev2 を用いて最適な条件を選択した．

4.2.1 パラフレーズの制限

自動で獲得したパラフレーズリストには間違ったパラフレーズも多く含まれており，それらを全てパラフレーズラティスに加えてラティスデコーディングするのは計算量も増大するため，1 文または 1 フレーズに対するパラフレーズの数を制限して実験を行った．パラフレーズ数は，1 フレーズあたりのパラフレーズ数を 3 個まで，1 文あたりのパラフレーズ数を文長（単語数）の 2 倍の数までに制限した．適用するパラフレーズを選ぶ基準としては，p，l，L の素性を用いる 3 通りの方法を試した．

4.2.2 素性の選択

パラフレーズラティス作成時の基準が p，l，L の 3 通りあるのに加えて，ラティスデコーディングで用いる素性の組み合わせとして，(p)，(p, l)，(p, L)，(p, l, d) の 4 通りあるため全部で $4 \times 3 = 12$ 通りの組み合わせがある．そこで，各組み合わせに対して dev1 でパラメータチューニングを行い，dev2 を用いて最も精度が良くなる組み合わせを選んだ．

4.3 結果

評価は BLEU によって行い，結果は，表 1 のようになった．これによると，英日翻訳では Moses に対して 1.36%，CCB に対して 1.10%，英中翻訳では Moses に対して 1.95%，CCB に対して 0.92%の BLEU 値の向上が得られた．精度は，Moses<CCB< 提案手法となることから，SMT にパラフレーズを組み込むことは有効であり，その手法としてパラフレーズによってフレーズテーブルを展開するよりもパラフレーズラティスに対してラティスデコーディングを適用するほうが効果が大きいといえる．

表 1: 実験結果 (%BLEU)

	Moses	CCB	提案手法
英日	38.98	39.24 (+0.26)	**40.34** (+1.36)
英中	25.11	26.14 (+1.03)	**27.06** (+1.95)

提案手法で選択された素性は，英日翻訳ではパラフレーズラティス作成時の基準は p，ラティスデコーディングの素性は (p, L) であった．また，英中翻訳ではパラフレーズラティス作成時の基準は L，ラティスデコーディングの素性は (p, l) が選ばれた．どちらも原言語の言語モデルが含まれた素性が選択されており，ラティスデコーディングの際に言語モデルを考慮することが有効であると言える．

5 まとめと今後の展開

本論文では，SMT において与えられた原文をパラフレーズによってパラフレーズラティスに変換し，それにラティスデコーディングを適用する手法を提案した．また，IWSLT2007 のデータセットを用いて評価実験を行ったところ，ベースラインである Moses に対して英日翻訳で 1.36%，英中翻訳で 1.95%の BLEU 値の向上が得られた．

今後の展開としては，単言語の巨大コーパスから自動で獲得したパラフレーズを用いて提案手法を適用することや，階層型フレーズベース SMT にパラフレーズを適用することを考えている．

参考文献

[1] Nicola Bertoldi, Richard Zens, and Marcello Federico. Speech translation by confusion network decoding. In *Proceedings of ICASSP 2007*, pp. 1297–1300, 2007.

[2] Christopher Dyer, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pp. 1012–1020, 2008.

[3] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of NAACL-2006*, pp. 17–24, 2006.

[4] Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. Improving statistical machine translation by paraphrasing the training data. In *Proceedings of IWSLT 2008*, pp. 150–157, 2008.

[5] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL-2005*, pp. 597–604, 2005.

[6] J Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL 2007*, pp. 967–975, 2007.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 177–180, 2007.