パターン翻訳と統計翻訳の結合

村上仁一　徳久雅人　池原悟
鳥取大学工学部知能情報工学科
発表者：サ　ミンソン

はじめに

- 機械翻訳の研究の歴史は大きく3 つの世代に分類できる

—第一世代：パターン翻訳

—第二世代：用例翻訳

—第三世代：統計翻訳

・現在，世界的にみると，統計翻訳に属する句ベース統計翻訳が主流である

つづき

- 統計翻訳には二つの問題点がある

①対訳データの量

ー翻訳モデルを学習するときに，対訳データが少量の場合，翻訳モデルの精度が低くなる．→翻訳精度が低くなる

②局所的な言語モデル

ー言語モデルには，通常N-gram モデルが利用されるが、局所的なモデルであり，文の構造を示す情報は少ない．そのため，非常に奇妙な文が出力される場合がある

- 本研究では，この2 点の問題点を解決するために，従来のパターン翻訳に統計翻訳を組み合わせた2 段階の翻訳を試みる．

システムの概要

- １．パターン翻訳

―始めに，入力の日本語に，パターン翻訳を利用して，英文を得る．

- ２．統計翻訳

―次に，統計翻訳を利用して，パターン翻訳から得られた英文を，英文に変換する．

予想されるシステムの利点（1）

1. 未知語

- 統計翻訳において，対訳データが少ないとき，未知語が多く出力される．（一般に人名地名などの固有名詞や数字）
- パターン翻訳を用いて，入力された日本語を英語に翻訳する　→統計翻訳の入力で未知語が少なくなる．

予想されるシステムの利点（2）

- 2．N-gramモデル

入力文に対して、

(a)適合するパターンがあっている場合

・得られる英文の語順は正しい

・N-gramは、局所的な言語モデルなので、翻訳精度が低くならない

(b)適合するパターンは間違っている場合

・得られる英文の語順は正しくない向上

・N-gramによってある程度補正される

実験データ

- 実験は日英翻訳のみとし，単文 と特許文の2 種類で行う.

文種別	学習データ	開発データ	テストデータ
単文	100,000	100	1,000
特許文	1,062,596	915	822

実験条件

- １．パターン翻訳：従来のパターンを利用した翻訳ソフトを利用する
- ２．統計翻訳：mosesを利用する
- ３．phrase tableの作成：mosesを利用する
- ４．言語モデル：SRILMを用いて学習する
  5-gramを用いる
- ５．パラメータチューニング：mosesを利用する
- ６．decorder：mosesを利用する

実験結果（単文）

- 入力文

  ーファイトがなくなってしまった。
- 正解文

  ーHe had little fight left in him .
- パターン翻訳

  ーThe fight has been lost.
- 提案手法

  ーThe struggle has gone.
- 統計翻訳mosesの出力

  ーファイトhas gone .

実験結果（特許文）

- **入力文**

ーこの場合、システム全体を制御するホストコンピュータ２６に専用線２９で接続されたプロセッサ２５に、クライオポンプ２１ａ、２１ｂの通信変換部２２ａ、２２ｂを通信ネットワーク２７で接続する。

- **正解文**

ーIn this case, communication conversion sections 22a, 22b of cryopumps 21a, 21b are connected by a communication network 27 to a processor 25 connected by an exclusive line 29 to a host computer 26 which controls the whole system.^.

- **パターン翻訳**

ーIn this case , the host computer connected to the communication network connection cryopump 21a , 21b and 22a , 22b communication lines 29 controls the whole system 26 in an exclusive processor 25 converter 27 .

- **提案手法**

ーIn this case , the transmission converter unit 22a and 22b of the cryopanels pumps 21a and 21b are connected to the processor 25 connected to the host computer 26 for controlling the entire system by an exclusive-use line 29 in a communication network 27 .

- **統計翻訳mosesの出力**

ーIn this case, the communication conversion parts 22a and 22b of the クライオpumps 21a and 21b are connected to the processor 25 connected to the host computer 26 which controls the whole system by the dedicated line 29 in the communication network 27.

評価

文種別		BLEU	NIST	Meteor
単文	moses	0.1260	4.3441	0.3654
	提案手法	0.1746	4.8083	0.4330
特許文	moses	0.2325	6.6049	0.5798
	提案手法	0.2953	7.3509	0.6313

まとめ

- 標準的な統計翻訳システムにおける問題点を解決するために，始めに，パターン翻訳を利用し，次に標準的な統計翻訳システムを利用することを考えた．実験の結果、BLEU やNIST やMeteor の値が向上し，提案した方式の有効性が示された．

- 今後，より最適な組み合わせを考えていきたい．