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
{murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

機械翻訳の研究の歴史は，コンピュータの誕生とともに始まっている．そして大きく 3 つの世代に分類できる．第一世代は”パターン翻訳”である．入力された日本語に対し，対応する日本語のパターンを検索する．次に日本語パターンに対応する英語パターンを得る．最後に英語パターンから英文を生成する．第二世代は”用例翻訳”である．あらかじめ，日本語と英語の対訳文を準備しておく．そして，入力された日本語文に対して，類似した日本語文を検索して，対応する英語文を出力する．最後に，この文を修正して出力文とする．第三世代は”統計翻訳”である．この方法は，言語を統計的に扱って翻訳する方法である．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．そしてツールが moses[13] としてソースつきで公開されている．そのため，対訳データがあれば簡単に統計翻訳の結果が得られる．

しかし，統計翻訳には多くの問題点がある．以下に本論文で着目する問題点を挙げる．

1. 対訳データの量
   統計翻訳には，翻訳モデルと言語モデルの 2 つのモデルが必要である．言語モデルは，対訳データである必要がないため，大量のデータが得られやすい．しかし，翻訳モデルを学習するときに，対訳データが少量のばあい，翻訳モデルの精度が低くなる．その結果，翻訳精度が低くなる．特に，翻訳されない単語が未知語として多く出力される．
2. 局所的な言語モデル
   言語モデルには，通常 N-gram モデルが利用される．しかし N-gram モデルは，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．この翻訳実験を試みた結果，精度の高い結果が得られた．

2　パターン翻訳と統計翻訳を組み合わせた翻訳

2.1　システムの概要

本研究では，対訳データの量と局所的な言語モデルの問題点を解決する方法として，従来のパターン翻訳と統計翻訳を組み合わせた，2 段階の翻訳を試みる．

具体的には，日英翻訳の場合を例に手順を示す．

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．

2.2　予想されるシステムの利点

パターン翻訳と統計翻訳を組み合わせた 2 段階による翻訳は，以下の利点が考えれられる．

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が未知語になりやすい．これらの名詞は，ルールにしたがって翻訳すれば，ある程度の高い翻訳が得られる．そこで，パターン翻訳をもちいて，入力された日本語を英語に翻訳することで，統計翻訳の入力において未知語が少なくなる．
   つまり，統計翻訳単独で翻訳するよりも，未知語が出力されなくなるため，翻訳精度が向上すると考えられる．これは特に対訳データが少ない場合，有効であると考えている．
2. N-gram モデル
   従来の統計翻訳では，言語モデルとして N-gram が利用されている．しかし N-gram は，局所的な言語モデルであり，大局的な言語モデルではない．特に，構文的な情報は，あまり持っていない．
   一方，従来のパターンベース翻訳は，大局的な言語モデルをパターンとして捕らえる．そして，入力された文に適応するパターンが合っていた場合，非常に精度の高い翻訳が得られる．しかし任意の文に対してパターンのカバー率が低いことや，複数のパターンが出力されるため，パターンの選択が困難であるなどの問題点がある．
   しかし，パターン翻訳を用いて英文を出力した後で，従来の統計翻訳をすることで，パターン翻訳が間違っていても翻訳精度は大きく減少しないと考えている．その根拠を以下に示す．
   （a）適合するパターンがあっている場合
   入力文に対して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と思われる．統計翻訳において使用される N-gram は，局所的な言語モデルであるため，英文の語順は正しい文は，翻訳精度が低くなることは，考えにくい．つまり，翻訳精度が高い文が得られる．
   （b）適合するパターンが間違っている場合
   入力文に対して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とは限らない．しかし，統計翻訳において使用される N-gram によってある

程度補正される．つまりパターン翻訳において翻訳精度が低い文は，統計翻訳によって，さらに翻訳精度が低くなるとは考えにくい．

したがって，全体として翻訳精度が向上する可能性が高い．

3 実験

3.1 実験データ

提案した，パターン翻訳と統計翻訳を組み合わせた 2 段階の翻訳の実験を行う．実験は日英翻訳のみとし，単文 [7][5] と特許文 [6][3] の 2 種類で行う．単文は，電子辞書から抽出した日本語が単文のデータベースである．英文は単文もしくは複文になっている．単文であるため短い文が多い．人間によるクリーニングが行われていて，信頼性は高い．特許文は ntcir07 で配布された文で，かなり長い文である．それぞれの実験にもちいた学習データと開発データとテストデータの量を表 1 に示す．

表 1　実験に使用したデータ量

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

3.2 実験条件

実験条件を以下に述べる．

1. パターン翻訳
   パターン翻訳には，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
   統計翻訳として moses[13] を利用する．
3. phrase table の作成
   翻訳モデルとして phrase table を利用する．phrase table は moses[13] にある train-factored-phrase-model.perl を利用して作成する．最大の phrase の単語数を規定する max-phrase-length は 20 とする．
4. 言語モデル
   言語モデルは SRILM[12] を用いて学習する．5-gram を用いる．スムージングとして -ukndiscount を利用する．
5. パラメータチューニング
   すべての実験には，reordering model を使用する．また moses[13] にある mert-moses.pl をもちいてパラメータチューニングをおこなう．
6. decorder
   decorder には moses[13] を利用する．decorder のパラメータは ttable-limit = 0 0 distortion-limit = -1 とする．その他は，default 値を利用する．

4 実験結果

4.1 単文

単文の実験で得られた例を以下に示す．表 2 は，入力した日本語文である．表 3 は，正解の英文である．表 4 は，表 2 を入力したときのパターン翻訳で得られた英文である．表 5 は，提案手法の英文出力である．表 6 は，標準的な統計翻訳である moses の英文出力である．

この表 5 と表 6 を比較すると，明らかに未知語が減少していることがわかる．

表 2　単文 入力文

1	ファイト が なくなっ て しまっ た 。
2	判別 困難 な 程度 の 出血 斑 が 認め られる 。
3	彼 は ぼうっと 空 の 一 点 を 見詰め て い た 。
4	会う たび に 彼女 は ますます 美しく なっ て いく 。
5	母 は わたし の 誕生 日 に チョコレート ケーキ を 焼い て くれ た 。
6	彼ら の うち 誰か 一 人 が 来 た 。
7	当社 で は 今年 は 女子 を 2 0 名 採用 し ます 。
8	手術 後 の 経過 は 順調 と いう 。

表 3　単文 正解文

1	He had little fight left in him .
2	Blood-spots that can be hardly identified are recognized .
3	He was blankly looking at a spot in the sky .
4	Each time I meet her , she is more beautiful .
5	Mother baked a chocolate cake for my birthday .
6	Someone of them came .
7	Our company will hire twenty women this year .
8	The operation lasted about three hours .

表 4　単文 パターン翻訳

1	The fight has been lost.
2	The bleeding spots of a grade with difficult distinction are accepted.
3	He was gazing at one point of empty vacantly.
4	She becomes still more beautiful whenever it meets.
5	The mother baked the chocolate cake on my birthday.
6	One someone came among them.
7	20 women are employed in our company this year.
8	It is said that the progress after an operation is favorable.

表 5　単文 出力 proposed

1	The struggle has gone .
2	The bleeding sites of the prize with difficult distinction are recognized .
3	He was gazing at one point of the sky absentmindedly .
4	She has grown more beautiful whenever you see .
5	The mother baked some chocolate cake for my birthday .
6	A someone came among them .
7	twenty women are employed in our company this year .
8	It is said that the progress after the operation is good .

表 6　単文 出力 baseline

1	ファイト has gone .
2	判別 difficult differing spot bleeding is recognized .
3	He had a 見詰め unsophisticated the sky .
4	I see her call on you have grown increasingly important .
5	My mother for my birthday cake chocolate .
6	Out of them one 誰か has come .
7	We provide a recruit 100 female athlete this year .
8	It was a fair way progress after the operation

4.2 特許文

特許文の実験結果を付録に示す．

表 8 は，入力した日本語文である．表 9 は，正解の英文である．表 10 は，表 8 を入力したときのパターン翻訳で得られた英文である．表 11 は，提案手法の英文出力である．表 12 は，標準的な統計翻訳である moses の英文出力である．

4.3 BLUE による評価

単文および特許文の BLEU[14] および NIST[14] および meteror[15] の評価を表 7 にまとめる．なお，参照文は 1 文である．表中の baseline は統計翻訳の moses を利用した実験結果である．また，表中の proposed は，パターン翻訳と統計翻訳を組み合わせた提案手法である．

この表から，提案手法 (proposed) が標準的な moses(baseline) より，いずれの値も向上していることがわかる．つまり，提案手法の有効性が認められる．

表 7　実験結果

文種別		BLEU	NIST	Meteor
単文	baseline	0.1260	4.3441	0.3654
単文	proposed	0.1746	4.8083	0.4330
特許文	baseline	0.2325	6.6049	0.5798
特許文	proposed	0.2953	7.3509	0.6313

5　考察

標準的な統計翻訳における言語モデルには，通常 N-gram モデルが利用される．しかし N-gram モデルは，局所的な言語統計モデルであり，文の構造を示す情報は少ない．そこで，文の構造を取り入れた言語モデルとして，多くの方式が提案されている．以下に例をあげる．

1. ネットワーク文法
   同じ形態素の位置が大きく異なる言語対に対して，翻訳する言語に合わせて，形態素を変えてから統計翻訳を行う方法が提案されている [8][9]．具体的には，日本語は，主語＋目的語＋動詞 (S+O+V) の語順である．一方，英語は主語＋動詞＋目的語 (S+V+O) である．そこで日本語をあらかじめ英語と同じ語順に変換し，それから従来の句ベースの統計翻訳を用いる．このようにすると，単語の翻訳の場所の移動が小さいため，翻訳精度の向上が期待できる．
2. 確率つき文脈自由文法
   言語モデルに確率つき文脈自由文法的な制約をいれることが試みられている [10]．しかし，文脈自由文法は N-gram と比較するとパラメータが非常に多くなるため，データが大量にあってもパラメータの値の信頼性が低い．そのため翻訳精度が向上しない場合もある．

今後，これらの言語モデルとの比較実験をおこなっていきたい．

6　まとめ

本研究では，標準的な統計翻訳システムにおける問題点である，未知語の問題と構文情報を利用していないことを解決するために，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、BLUE 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．

この方式は，前段のパターン翻訳の翻訳精度によって，全体の翻訳精度がある程度決まると思われる．今後，より最適な組み合わせを考えていきたい．

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, “The Mathematics of Statistical Machine Translation, Parameter Estimation”, Computational Linguistics, 19(2), 1993.

[2] Jin’ichi Murakami, Masato Tokuhisa, Satoru Ikehara, “Statistical Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables”, International Workshop on Spoken Language Translation 2007, pp.151-155, 2007.

[3] 内山将夫, 山本幹雄, 藤井敦, 宇津呂武仁, “特許情報を対象とした機械翻訳 -共通基盤による評価タスクを目指して-”, 情報処理学会研究報告, Vol. 2007, No.(2007-NL-180), pp.133-138, 2007.

[4] Philipp Koehn, Franz J. Och, and Daniel Marcu, “Statistical phrase-based translation”, In Proceedings of HLT-NAACL 2003, pp. 127-133, 2003.

[5] 西山七絵, 村上仁一, 徳久雅人, 池原悟, “単文文型パターン辞書の構築”, 言語処理学会第 11 回年次大会, pp.372-375, 2005.

[6] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, ” Overview of the Patent Translation Task at the NTCIR-7 Workshop”, Proceedings of the 7th NTCIR Workshop Meeting, 2008.

[7] 村上仁一, 池原悟, 徳久雅人, “日本語英語の文対応の対訳データベースの作成”, 「言語，認識，表現」第 7 回年次研究会, 2002.

[8] Yushi Xu, Stephanie Seneff, ”Two-Stage Translation: A Combined Linguistic and Statistical Machine Translation Framework”, Proceedings of the Eighth Conference of the Association for Machine Translation (AMTA) 2008.

[9] Jason Katz-Brown,Michael Collins, ” Syntactic Reordering in Preprocessing for Japanese English Translation: MIT System Description for NTCIR-7 Patent Translation Task”, Proceedings of the 7th NTCIR Workshop Meeting, 2008.

[10] Katsuhito Sudoh, Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki, ”NTT Statistical Machine Translation System for IWSLT 2008”, IWSLT 2008, pp.92-97 (2008).

[11] GIZA++, http://www.fjoch.com/GIZA++

[12] SRILM, The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm

[13] Moses, moses.2007-05-29.tgz, http://www.statmt.org/moses/

[14] NIST Open Machine Translation, http://www.nist.gov/speech/tests/mt

[15] The METEOR Automatic Machine Translation Evaluation System, http://www.cs.cmu.edu/ alavie/METEOR/

表 8　特許文 入力文

1	さらに 、 図 4 に 示す よう に 、 システム 全体 を 制御 する ホスト コンピュータ 4 6 も 通信 ネットワーク 4 7 上 に 接続 する こと によって 、 ホスト コンピュータ 4 6 と プロセッサ 4 4 と を 接続 する 専用 線 を なくす こと が できる 。
2	この 場合 、 システム 全体 を 制御 する ホスト コンピュータ 2 6 に 専用 線 2 9 で 接続 さ れ た プロセッサ 2 5 に 、 クライオポンプ 2 1 a 、 2 1 b の 通信 変換 部 2 2 a 、 2 2 b を 通信 ネットワーク 2 7 で 接続 する 。
3	また 、 圧縮 機 ユニット 3 3 に も 、 I ／ O 変換 部 （ 図示 せ ず ） が 搭載 さ れ て い る 。
4	これ に対して 、 I D コード 格納 部 9 3 に 格納 さ れ て いる I D は 、 当該 クライオポンプ 3 1 自身 に 割り当て られ て いる I D で ある 。
5	そして 、 通信 変換 部 3 2 によって 、 受け取っ た 信号 が 伝搬 に 適し た 信号 フォーマット に 変換 さ れ 、 通信 ネットワーク 3 7 に 送信 さ れる 。
6	ステップ S 1 1 で 、 操作 端末 器 6 0 から 割り込み 入力 が ある か 否 か が 判別 さ れる 。
7	ステップ S 3 で 、 当該 パケット の 通信 データ が 読み込ま れ て 、 I ／ O 変換 部 7 6 に 渡さ れる 。

表 9　特許文 正解文

1	Further, by connecting the host computer 46, which controls the whole system, also onto the communication network 47 as shown in FIG. 4, the exclusive line for connecting the host computer 46 and the processor 44 to each other can be eliminated.
2	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.
3	An I/O conversion section (not shown) is mounted also on the compressor unit 33.
4	In contrast to this, the ID stored in the ID code storage section 93 is an ID assigned to the relevant cryopump 31 itself.
5	Then, the communication conversion section 32 converts the received signal into a signal format suited to propagation, and transmits the signal to the communication network 37.
6	At step S11, it is decided whether or not an interrupt input from the operation terminal unit 60 is present.
7	At step S3, communication data of the packet is read and delivered to the I/O conversion section 76.

表 10　特許文 パターン翻訳

1	Furthermore, as shown in Fig. 4, the dedicated line which connects a host computer 46 and a processor 44 can be lost by connecting the host computer 46 which controls the whole system on the communication network 47.
2	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.
3	Moreover, the I/O conversion part (not shown) is carried also in the compressor unit 33.
4	On the other hand, ID stored in the ID code storing part 93 is ID currently assigned to the クライオ pump 31 self concerned.
5	And the received signal is changed into the signal format suitable for propagation by the communication conversion part 32, and is transmitted to the communication network 37.
6	It is distinguished whether at Step S11, it interrupts from the operation terminal machine 60, and there is any input.
7	At Step S3, the communication data of the packet concerned is read and the I/O conversion part 76 is passed.

表 11　特許文 出力 proposed

1	As shown in FIG . 4 , a private line which connects the host computer 46 to the processor 44 can be eliminated by connecting the host computer 46 for controlling the entire system on the network 47 .
2	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 .
3	The I / O converter (not shown) is installed in the compressor unit 33 .
4	On the other hand , the ID stored in the ID code storing portion 93 is ID allocated to the cryopump 31 itself .
5	The received signal is converted into a signal format which is suitable for transmission by the transmission converting unit 32 , and then transmitted to the communication network 19 .
6	It is determined whether or not in step S11 , the interrupt from the console unit 60 , and there is a key input .
7	In step S3 , the communication data of a packet is read to the I / O converter 76 is provided .

表 12　特許文 出力 baseline

1	As shown in FIG . 4 , the host computer communication network 46 or 47 on the host computer 44 and 46 are connected in a processor for controlling the entire system by connecting the private line can be eliminated .
2	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 .
3	The compressor unit 33 in the I / O converter (not shown) are mounted .
4	On the other hand , the ID code stored in the ID storage means 9 , the cryopump 31 ID assigned to itself .
5	The communication unit 32 , the received signal is converted into a signal format which is suitable for a communication network 37 is transmitted to the transmission data .
6	At step S11 , it is determined whether or not there is an input signal from a terminal control device 60 .
7	In step S3 , the communication of the packet data is read and transferred to the I / O converter 76 .