フレーズ拡張したワードラティスを用いた対訳コーパスのない言語からの統計的機械翻訳

楠本 高康　秋葉 友良
豊橋技術科学大学
{kusumoto, akiba}@cl.ics.tut.ac.jp

1　はじめに

インターネットの発達などにより，外国語の情報にアクセスする機会が増えた．しかし，外国語の翻訳には人的・時間的なコストがかかるため，有益と思われる情報すべてを翻訳することはできない．そのため，人的なコストを掛けずに文章を翻訳できる，機械翻訳の技術がますます重要になっている．

統計的機械翻訳は，大量の対訳文 (対訳コーパス) を分析し，対訳文同士の統計量を分析することで，翻訳の規則を学習する手法である．近年，計算機が進歩し対訳コーパスが整備されるにつれ，統計的機械翻訳の性能が向上した．統計的機械翻訳には (1) 翻訳規則の記述に専門知識を持った人間を必要としない．(2) 対訳コーパスがあるあらゆる言語間で機械翻訳をすることができる．という利点がある．

対訳コーパスには，翻訳したい言語ペアの対訳コーパスが，必ずしも利用出来るとは限らず，また対訳コーパスを新規に作成する作業は非常に労力がかかる．対訳コーパスがない言語ペア (原言語-目的言語) 間で統計的機械翻訳をする方法として，中間言語を用いる方法が提案されている [3]．これは，原言語と目的言語両方との間に対訳コーパスを持つ，中間言語の存在を仮定し，中間言語を経由することで，原言語を目的言語に翻訳する手法である．様々な言語との間に対訳コーパスを持つ英語が，しばしば中間言語として利用される．しかしこの方法は，原言語と目的言語の両方について中間言語との対訳コーパスが必要なので，原言語と中間言語，あるいは中間言語と目的言語，いずれかの対訳コーパスが利用できない場合は適用できない．

本研究では，中間言語 (英語) と目的言語 (日本語) の間にのみ対訳コーパスがある場合に，統計的機械翻訳手法を適用する方法を提案する．そのような言語ペアとして，ベトナム語-日本語間の翻訳を扱う．

ベトナム語を英語に翻訳するために，ベトナム語-英語単語辞書を利用する．この翻訳には，辞書による候補のうちどの英単語を選ぶか，英単語をどのように並べるか，単語辞書に含まれない語をどのように補完するかという問題がある．Mahn ら [2] は，英単語の選択と語順の問題を解決するために，英文ラティスを用いる方法を提案した．ラティスとは文の複数候補を効率良く表現する形式である．音声認識結果の複数候補からの翻訳を目的に，ラティス表現された文を入力とする統計的機械翻訳のラティスデコーダが提案されている [4]．Mahn らは，ラティスを用いて単語辞書の複数候補を表現し，英日フレーズテーブルに含まれるフレーズを英文ラティスに追加することで語順の訂正を行った．

単語の関係などを表す役割を持つ機能語は，文章に大きな意味を持つ．しかし，機能語は，異なる言語間で1対1に対応しているわけではないため，英文には必要だがベトナム語-英語単語辞書では翻訳できない機能語が存在する．本研究では，作成した英文ラティスに，Mahn が提案した語順の訂正を行うとともに，機能語の補完を行うことで翻訳の改善を試みた．

2　英文ラティスの作成

ベトナム語-英語単語辞書を用いて，ベトナム語を英文ラティスに翻訳する方法を示す．本研究では，有限状態変換器ライブラリ OpenFST[1] を用いて，ラティスの作成及び拡張をした．

英文ラティスの作成方法を，例とともに以下に示す．

1. ベトナム語文が入力として与えられる
   『Kinh t th gii ang khng hong ti chnh』
2. セグメンテーションツールを用いて，ベトナム語文を単語に分割する
   『Kinh t | th gii | ang | khng hong | ti chnh』

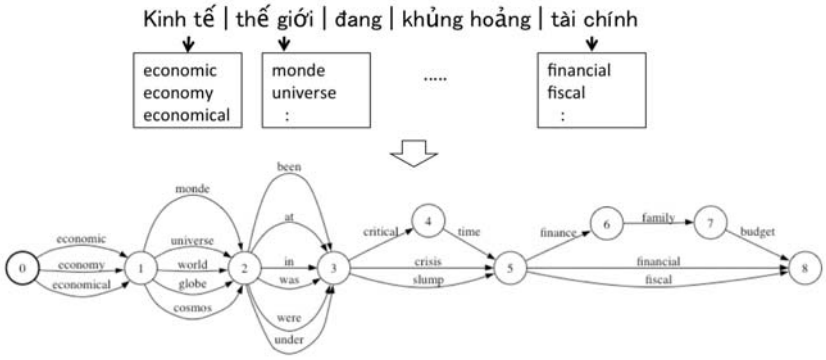

図 1: 英文ラティスの作成

3. ベトナム語-英語単語辞書を用いて，ベトナム語単語を英単語に翻訳する

ベトナム語	翻訳英単語 (候補)
Kinh t	economic, economy, economical
th gii	monde,universe, world, globe, ...
ang	were, under, been, at, in, was
khng hong	critical time, crisis, slump
ti chnh	financial, fiscal, ...

4. 手順 3 で求めた英単語の候補と，もとのベトナム語単語の位置に基づいて，英文ラティスを作る．例文からは，図 1 のようなラティスができる．

以上の手順を踏むことで，ベトナム語文を英文ラティスに変換することができる．各ベトナム語単語の翻訳語は，元のベトナム語単語と同じ位置になる．

3 ラティスの拡張

2 節で作成した英文ラティスは，語順がベトナム語のものであるため，英語の文法になるよう並び替えなければならない．また，単語辞書には英文に必要な機能語が含まれていない場合があるため，英文ラティスに欠けている機能語を補わなければならない．そこで，英文ラティスのデコードに用いる英日統計的機械翻訳システムの，英日フレーズテーブルに含まれるフレーズを，英文ラティスに追加することでそれらの問題を解決した．フレーズテーブルに含まれるフレーズは，英文ラティスを日本語にデコードする際に採用される可能性が高く，有用な候補になる．

3.1 語順の訂正

ベトナム語と英語の文法は，共に S-V-O の形であり類似しているため，英文ラティスの文法は，おおむね正しい．しかし，名詞句の語順 (修飾語と被修飾語の順序など) など，異なっているものもあるため，それらを英文法通りに訂正する必要がある．

英日フレーズテーブルに含まれているフレーズは，正しい英語の語順を反映している．また，フレーズテーブルに含まれるフレーズは必ず対応する日本語翻訳を持つため，デコード中に候補フレーズとして採用される可能性が高い．そこで，英文ラティス中のあるパスに対応する英単語列を並び替えたフレーズが，英日フレーズテーブルに含まれていた場合，元のパスと並列に，そのフレーズを新たなパスとしてラティスに追加を行う．言い換えると，フレーズテーブルを事例ベースとして用いて，フレーズ単位で語順の訂正候補をラティスに追加することで，語順の訂正を行う．

この手順を，例と共に以下に示す．

1. あらかじめ，フレーズテーブルの内容を，以下のルールでハッシュテーブルに保存しておく．

   キー フレーズの単語を辞書順に並び替えたもの

   キーの値 辞書順に並び替える前のフレーズ

表 1: ハッシュの中身（一部）

ハッシュのキー	ハッシュの値
economy in world	“world economy in”, “economy...
crisis financial in	“in crisis financial”, “financial...
...	...

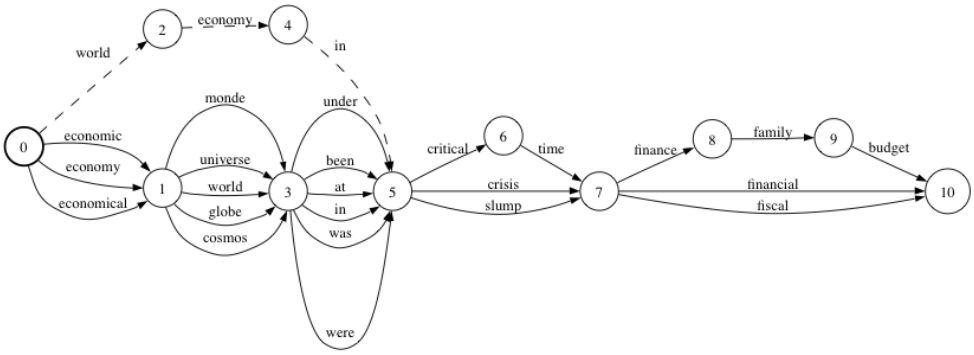

図 2: 語順の訂正（破線部は追加したパス）

2. 現在のノードからの N 単語パスを探索する．探索により得られた N 単語の部分単語列を W_N とする．探索の途中でラティスの終端ノードに到達した場合は W_N は空文字とする．

3. W_N を辞書順に並び替えたものをキーとして，予め作成しておいたハッシュを引く．ハッシュの返戻値は，W_N を並び替えてできる英日フレーズテーブルに含まれるフレーズ集合 V である．キーに該当する値がなければ空集合が返されるとする．

4. W_N の開始ノードと，終了ノードの間に，V に含まれる各フレーズを表すパスを追加する．

5. ラティスにパスを追加することで非決定性状態遷移グラフになった場合，ラティスを決定化する．

6. すべてのノードにおいて，手順 2 から 5 を繰り返す．

以上の手順を踏むことで，フレーズテーブルに含まれるフレーズをラティスに追加することができる．図 1 のラティスに，英日フレーズテーブルに含まれるフレーズ『world economy in』を追加し，語順の訂正を行った例を図 2 に示す．

3.2 機能語の補完

作成した英文ラティスに機能語を追加するために，ハッシュテーブルを以下のように拡張した．

ハッシュのキー フレーズの単語を辞書順に並び替えたもの

ハッシュのキー 2 フレーズの単語を辞書順に並び替えたものから前述の機能語を取り除いた物

ハッシュの値 辞書順に並び替える前のフレーズ

ラティスに追加する機能語として，英文新聞記事 20 万文に頻出する機能語の中で，“not” のように，追加すると文の意味が変わってしまう単語を除いた in, to, of, the, on, for, is, that, by, a, an を選んだ．ハッシュの例を表 2 に示す．

表 2: ハッシュの中身（機能語拡張）

ハッシュのキー	ハッシュの値
economy in world	“world economy in”, ...
economy world	“world economy in”, ...
crisis financial in	“in crisis financial”, ...
crisis financial	“in crisis financial”, ...
...	...

このハッシュを用いて，3.1 節の同様の操作を行う．W_N が機能語を含んでいなくても，W_N に機能語を含めたハッシュのキーに相当するため，ラティスに機能語を補完することができる．

4 実験

4.1 実験に用いたリソース

本実験では，ベトナム語-英語単語辞書として，Free Vietnamese Dictionary Project[1] のベトナム語-英語辞書と英語-ベトナム語辞書を統合して使用した．単語辞書のベトナム語見出しは 137,949 語，英語の翻訳は 271,540 語である．ベトナム語のセグメンテーションツールとしては，vnTokanizer [2] を用いた．

[1] http://tudientiengviet.net/data.html
[2] http://www.loria.fr/ lehong/tools/vnToolkit.php

フレーズ拡張	並び替え	並び替え+機能語追加	追加フレーズ数
拡張なし	0.1570	0.1570	0
N=2	0.1608	0.1631	10824
N=3	0.1572	0.1597	1668
N=4	0.1550	0.1546	104
N=5+4	0.1546	0.1538	106

表 3: 翻訳結果の比較 (BLEU)

英日機械翻訳の対訳コーパスとしては，読売新聞 1999 年度-2001 年度の新聞記事の英日対訳文 150,000 文ペア [5] を用いた．そのうち 200 文ペアをテスト及びチューニングに用い，残りの 148,000 文ペアを学習データとして使用した．また，日本語の言語モデルとして，対訳コーパスの日本語側 148,000 文の 3-gram モデルを用いた．

抽出した対訳文 200 文ペアのうち，98 文ペアをテストセットとして用いた．テストセット 98 文ペアの英文側を，ベトナム語話者がベトナム語に翻訳した 98 文をテストセットとして用い，残りの 102 ペアを英日統計的機械翻訳のチューニングに用いた．

4.2 実験結果

単語辞書を使って作成した英文ラティス，語の並び替えを行ったもの，語の並び替え及び機能語の補完をしたものそれぞれについて，追加するフレーズの単語数 N を変化させながら，越日統計的機械翻訳の翻訳文の 2-gram の BLEU を調べた．実験結果を表 3 に示す．表の列は，左から順に，語順の訂正のみ行った時の BLEU，語順の訂正及び機能語を追加したときの BLEU，98 個の英文ラティスに追加したフレーズの総数である．表の行は，ラティスに追加するフレーズの単語数であり，最上段はラティスの拡張を行わなかった場合，最下段は N が 5 または 4 であるフレーズを追加した場合である．

N=2 の時もっとも多くのフレーズが追加されており，1 文当たり平均して 100 フレーズ以上が追加されている．N=3 のときは，語順を入れ替えたものよりも，それに加えて機能語の訂正を行ったものの方が BLEU が大きくなった．一方，N=4,N=4+5 の場合は，語順の訂正のみを行った場合のほうが BLEU が大きく，またフレーズ拡張をしない場合よりも BLEU が低くなった．これは，統計的機械翻訳において，語長が大きいフレーズは翻訳の手がかりとして重要になる一方で，ラティスを並び替えた際に偶然合致するフレーズができた場合に誤って採用された場合の悪影響が大きいからだと思われる．

5 おわりに

本実験では，中間言語と目的言語の間にのみ対訳コーパスが存在する言語間において，統計的機械翻訳を行う手法を提案した．

単語辞書に含まれない単語の処理をどうするか，またラティスの作成を高速に行うため，見込みのないパスを枝刈りする方法などが今後の課題である．

参考文献

[1] Cyril Allauzen et al. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of Twelfth International Conference on Implementation and Application of Automata, (CIAA 2007)*, pp. 11–23, 2007.

[2] Nguyen Manh Hung, 秋葉友良. Word lattice decoding を利用した対訳コーパスがない言語からの統計的機械翻訳. 言語処理学会第 16 回年次大会講演論文集, pp. 1006–1009, 2010.

[3] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based stastical machine translation. In *Proceedings of NAACL / HLT*, pp. 484–491, 2007.

[4] R. Zhang, G. Kikui, H. Yamamoto, and W. Lo. A decoding algorithm for word lattice translation in speech translation. In *Proceedings of the International Workshop on Spoken Language Translation*, 2005.

[5] 内山将夫, 井佐原均. 日英新聞の記事および文を対応付けるための高信頼度尺度. 自然言語処理, Vol. 10, No. 4, pp. 201–220, 2003.