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

Nguyen Manh Hung　　秋葉友良

豊橋技術科学大学

1　はじめに

統計的機械翻訳は、ある言語ペア（ソース言語とターゲット言語）の対訳コーパスに基づいて学習した翻訳規則に基づき翻訳を行う手法である。大量の対訳コーパスがあれば、人手での翻訳規則の構築なしに、安価に翻訳システムが構築できる手法として有望である。普及した言語ペア、例えば欧州の各国の間[1]や英語-日本語などに対しては対訳コーパスが大量に存在する。一方、対訳コーパスが少ないか利用できない言語ペア、例えばベトナム語`-日本語や日本語-フランス語など、も多い。対訳コーパスのない言語ペアでは、統計的機械翻訳をそのまま通用することは困難である。

対訳コーパスのない言語ペアに対する手法として、中間言語を利用する手法が提案されている。この手法では、ソース言語と中間言語の間、および中間言語とターゲット言語の間、それぞれについて対訳コーパスが利用できると仮定する。そして、各言語ペアから学習した２つの翻訳システムを順に通用して翻訳するか、学習した２つのフレーズテーブルを結合して翻訳を行う[3]。しかし、中間言語を介して、２つの対訳コーパスが利用できない言語対に対してはこの方法は利用できない。

本研究ではターゲット言語（日本語）と中間言語（英語）の間にのみ対訳コーパスが存在する場合の統計的機械翻訳の手法を提案する。そのような言語ペアの例として、ベトナム語から日本語への翻訳に焦点をあてる。ベトナム語は日本語との間に対訳コーパスが存在しない上に、利用可能な機械可読辞書も少ない。また日本語とベトナム語では語順が大きく異なる。ベトナム語の語順は S-V-O 型であるが、日本語は S-O-V 型である。

提案方法は、英語を中間言語として用い、ベトナム語-英語の単語辞書と英語-日本語の対訳コーパスを用いて統計的機械翻訳を実現する。まず、入力ベトナム語文を、ベトナム語-英語の単語辞書を用いて、英語ラティスへ変換する。ラティスは複数文候補の表現形式であり、辞書による翻訳が多義である場合でも効率よく表現できる。次に、英語-日本語の対訳コーパスから学習した英日統計的機械翻訳によって、英語ラティスを日本語文に翻訳する。その際、ベトナム語と英語の語順の差異に対応するため、英日翻訳に用いるフレーズテーブルを参照しながら、英語ラティス中の単語の並び替えを行い、新たなパスとして英語ラティスに追加する。英語ラティスからの翻訳には、ラティスデコーダを用いる。

評価実験の結果、関連する研究に比べて有望な結果が得られた。

[1]http://www.statmt.org/europarl/

2　連関研究

ベトナム語―日本語の機械翻訳に関する研究は少ない。ルールベース翻訳システムの研究としては、My Chau[2]らの研究がある。一方、統計的機械翻訳の研究にはTuanら[1]の研究がある。Tuanらは、英日対訳コーパスから越日対訳コーパスを自動生成して、統計的機械翻訳を行う手法を提案している。英語―ベトナム語の変換方法は、英語―ベトナム語の単語辞書によって語ベース変換を行い、ベトナム語の言語モデルを用いて確率の高い１Bestのベトナム語文を選ぶという方法である。

統計的機械翻訳において、対訳コーパスが利用できない言語対に対して、中間言語を利用する手法が提案されている。これらの手法では、ソース言語と中間言語の間、及び中間言語とターゲット言語の間、それぞれについて対訳コーパスが存在すると仮定する。直接的な方法は、対訳コーパス

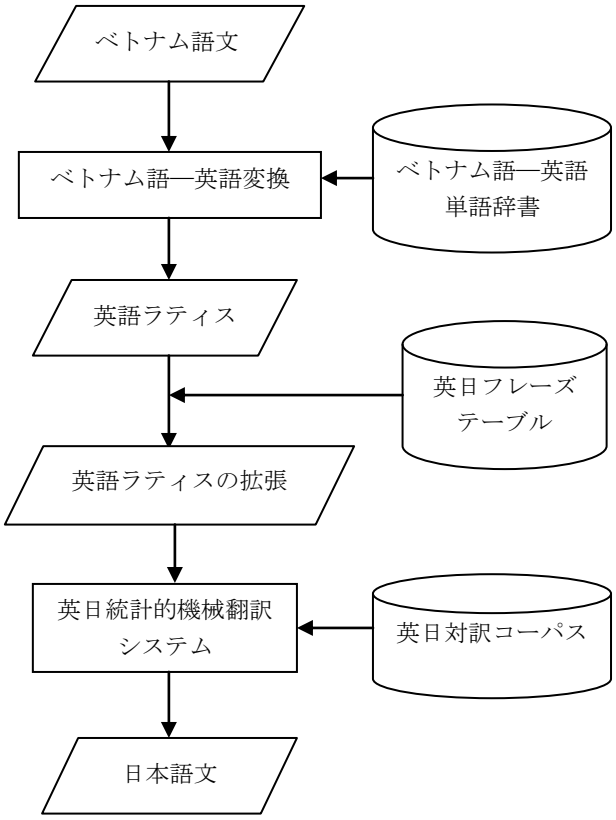

図１：システムの流れ

から２つの統計的機械翻訳システムを構築して２段階に翻訳を行う方法である。Utiyamaら[3]は、対訳コーパスで学習した２つの翻訳モデルを統合して、１段階で翻訳を行う手法を示している。WuとWang[4]の研究では、翻訳精度を改善するため、複数の中間言語を同時に利用している。

ラティスデコーダは、本来は、複数の入力文（認識候補）を扱う必要のある音声翻訳システムのために導入された[10]。音声入力以外に適用した例として、Dyerら[5]はアラビア語から英語、および中国語から英語への翻訳において、ソース言語（アラビア語、中国語）の形態素分割の多義性をラティスで表現することにより、翻訳性能を改善したと報告している。

3　提案方法

提案手法における翻訳の手順を以下に示す。（図１）

(I)　ベトナム語の文を、ベトナム語-英語の単語辞書に基づいて、英語ラティスへ変換する。

(II)　英日のフレーズテーブルを参照してラティスを拡張し、語順の候補を増やす。

(III)　生成した英語のラティスを英日 SMT システムで翻訳する。

以下では、これらのステップを順に説明する。

3.1　英語ラティスの生成

ラティス表現は複数の入力候補を表現するために利用される。ラティスは有向非循環グラフであり、エッジには単語が与えられる。開始ノードから終了ノードまでの１つのパスが、１つの入力文を表している。以下では、入力ベトナム語文から英語ラティスを生成する手順を例とともに示す。

(0) 入力ベトナム語例文「英語」：

Kinh tế thế giới đang khủng hoảng tài chính
「World economy is in financial crisis」

(1) ベトナム語文をセゲメンテーションツール[(2)]で単語に分割する。文頭、文末および分割された単語間にラティスのノードを生成する。

Kinh tế | thế giới | đang | khủng hoảng | tài chính

(2) 分割されたベトナム語単語から単語辞書を引き、訳語の候補を取り出す。

kinh tế: economic, economy, economical
thế giới : monde, universe, world, globe, cosmos
đang : were, under, been, at, in, was
khủng hoảng : critical time, crisis, slump
tài chính : financial, fiscal,...

(3) 各訳語候補について、対応するラティスのノード間にエッジまたはパスを生成する。訳語が１単語の場合はその単語のラベルを持つエッジを、複数単語から成る場合には単語エッジ間にノードを置いたパスを生成する。例文からは、図２のラティスが生成される。

3.2　ラティスの拡張

ベトナム語と英語は、共に S-V-O 型で文法は類似しているが、名詞句の語順が異なるなど、必ずしも語順は一致しない。したがって、作成したラティスの語順はまだ十分に正確ではない。そこで、日英対訳コーパスから学習したフレーズテーブル

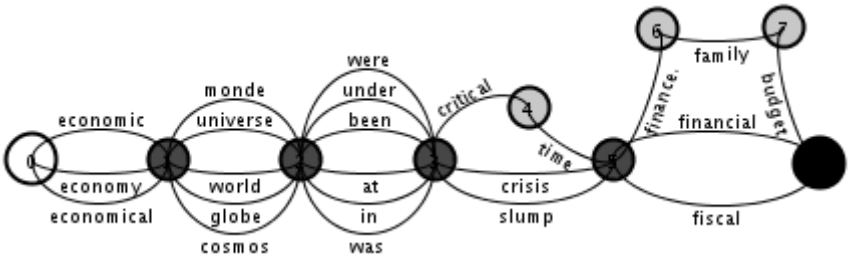

図２：ラティスの例

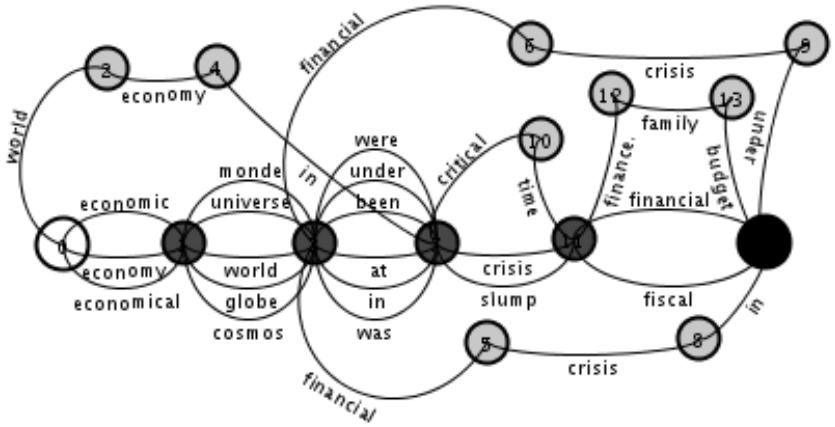

図３：ラティス拡張の例

[7]を参照して、語順を並び替えた候補をラティスに追加する。フレーズテーブルは、英日対訳コーパスから学習したフレーズを保管するので、載っている英語のフレーズは正しい語順である可能性が高い。またフレーズテーブルから選択したフレーズは、デコードの時に選択される可能性が高く、翻訳に取って有用な候補でもある。

ラティスの拡張は次の手順で行う。英語ラティス中のすべての N 単語パス（始端ノード S、終端ノード E とする）について、その N 単語を並び替えたフレーズがフレーズテーブル中に存在するかどうか調べる。存在する場合には、フレーズテーブルに掲載されている語順の新たな N 単語パスを、ノード S からノード E の間に追加する。

例文に対してフレーズ拡張を作った結果を図３に示す。ここでは、図２のラティスにある３単語のパス「economy - world - in」について、フレーズテーブルに異なる語順「world – economy - in」が見つかり、それを追加した。他のフレーズ「financial – crisis – in」等についても同様である。

[2]http://www.loria.fr/~lehong/tools/vnToolkit.php

[3]http://tudientiengviet.net/data.html

4　実験

4.1　データ

ベトナム語―英語の辞書は Free Vietnamese Dictionary Project[(3)]に存在するものを使用した。項目数は約 9.5 万語である。

英日対訳コーパスは読売新聞 1999-2001 年度の新聞記事に対して対訳関係を求めた対訳コーパス[6]を利用した。サイズは 150,000 文ペアである。その内、200 文ペアをテスト用に抽出し、148,800 文ペアを学習データとした。言語モデルの学習には、対訳コーパスの日本語側を用いた。

ベトナム語―日本語テストセットはテスト用に抽出した英日テストセットから作成する。英日のテストセットの英語部分から人手によって日本語部分も参考にしながらベトナム語へ翻訳した。

言語モデル生成ツールには SRILM[9]を、デコーダには Moses[8]を利用した。評価指標には、正解訳に対する３次の BLEU スコアを用いた。

4.2　翻訳結果

実験の目的は以下の２つである。一つ目はベトナム語―日本語の統計的機械翻訳が対訳コーパスが存在しなくても可能であることを示すことである。二つ目は提案した方法の内、どの手法か最も効果があるかを調べることである。

実験１．提案法の効果

表１：提案手法の効果

手法	Bleu スコア
Baseline	5.42
Lattice	11.72
Lattice + PTb	**12.17**
英日 SMT(上限)	32.13

提案手法によるベトナム語-日本語の翻訳性能を調べた。ベースライン手法として、Tuan らの手法と比較した。提案手法としてラティスデコーダを用いる手法（Lattice）とフレーズテーブルをつかってラティスの拡張を行う手法(Lattice+PTb)との比較を行った。結果を表１に示す。提案した手法はベースラインを超えて Bleu スコアの 6.75 点で改善できた。

実験２：ラティス拡張の効果

フレーズテーブルを参照するフレーズ長 N を変化させて、どのぐらいが最も効果が得られるかを調査した。ここで N4、N5、N6、N45 はそれぞれ、フレーズ長４、５、６、フレーズ長４と５両方を利用した場合、である。また各手法において、拡張できたフレーズ数を調べた。結果を表２に示す。

拡張できたフレーズ数が多ければ多いほど Bleu スコアが向上しており、フレーズ拡張の効果が示されている。N6 の場合はフレーズ拡張できず、拡張しない場合と同じ値となった。長さ４と５を両方用いた場合、参照数が最も多くなり、最も良い翻訳性能を示した。

表２：ラティス拡張の比較

N グラム	Bleu スコア	参照フレーズ数
Lattice	11.72	
Lattice + N4	11.929	98
Lattice + N5	11.82	10
Lattice + N45	12.17	108
Lattice + N6	11.72	0

6 まとめ

本研究では、対訳コーパスがない言語ペアの統計的機械翻訳手法を提案し、評価実験によりその効果を示した。本手法は、ベトナム語-日本語ペアだけでなく、様々な言語ペアへ適用することができると考えている。今後の課題として、異なる言語ペア（フランス語―日本語、ベトナム語―欧州の国の言語など）で提案手法を評価する。また、ベトナム語―日本語の翻訳精度を向上させるため、英語以外の中間言語を利用できるかどうか調査したい。

参考文献

[1] Le Tuan Anh, 秋葉友良. パラレルテキストの自動生成に基づく越日統計的機械翻訳. 言語処理学会第 14 回年次大会, pages 997-1000. 2008.

[2] 田中友樹, Nguyen My Chau, 池田尚志. 日本語―ベトナム語機械翻訳システム jaw/Vietnamese における翻訳実験－連体修飾（「の/こと」など)、86 例文に対する実験. 言語処理学会第 13 回年次大会, pages 674-677. 2007.

[3] Masao Utiyama, Hitoshi Isahara. A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation. In *Proc. of NAACL HLT 2007*, pages 484–491. 2007.

[4] Hua Wu, Haifeng Wang. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proc. of ACL-07*, pages 856–863. 2007.

[5] Christopher Dyer, Smaranda Muresan, Philip Resnik. Generalizing Word Lattice Translation. In *Proc. of ACL-08: HLT*, pages 1012–1020. 2008.

[6] 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性尺度. 自然言語処理, 10(4), pages 201-220. 2003.

[7] Philipp Koehn, Franz Josef, Och,Daniel Marcu. Statistical Phrase-Based Translation. In *Proc.of HLT-NAAC- 2003*, pages 127-133. 2003.

[8] P. Koehn, H. Hoang, M. Federico, N. Bertoldi and others. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL -07*, pages 177–180. 2007.

[9] A. Stolcke. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the ICSLP*, pages 901-904. 2002.

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