単言語または二言語の分割性による類推翻訳の検討

西川 裕介　木村 竜矢　松岡 仁　ルパージュ・イヴ
早稲田大学大学院 情報生産システム研究科
y_nishikawa@asagi.waseda.jp, tatsuya-kimura@ruri.waseda.jp, jinmatsuoka@akane.waseda.jp, yves.lepage@waseda.jp

1　はじめに

機械翻訳を行う上で問題となるのが言語間での文法の違いである。これを解決するため様々な手法が研究されている。本論文は Lepage ら [5] によって提案された類推関係に基づく用例翻訳 (類推翻訳) を対象としている。本論文での類推関係とは A:B::C:D と表現され、A と B の関係は C と D の関係に等しいことを意味する。これにより A、B、C がわかっていれば翻訳テーブルに存在しない D も導くことができる。ただし、この導出も参照する翻訳テーブルの影響を大きく受ける。翻訳テーブルは Anymalign[1] [3] や GIZA++[2] [6] のような従来手法で作成できる。しかし、これらは統計的機械翻訳 (Statistical Machine Translation, SMT) 向けのツールであり用例翻訳ではそのような手法は確立されていない。本論文では類推翻訳システムを対象とした翻訳テーブルを作成する手法を検討する。本論文ではその手法として可切性 (Secability) を用いる。また、可切性は二言語の文章を独立して構造化する手法であるため二言語間で並列的に構造化する手法と比較する。そのためにそれぞれの手法で作成したテーブルを使った翻訳実験を行い BLEU 値によって評価する。

2　関連研究

可切性は Chenon[1] によって提案された 。バイリンガルコーパスの文節などに統計的根拠を与えることで翻訳メモリをより有効に活用するために木構造解析を行い、この木構造モデリングを定式化した。また、竹谷 (2012) によって用例翻訳システムを対象とした翻訳テーブルの作成からその有効性を検討する研究が行われた [7]。

類推翻訳システムは Lepage ら [5] によって提案された。彼らの用例翻訳システムは運用するにあたってより長い用例を参照することで BLEU の評価値が向上した。本研究ではこの結果をもとに翻訳テーブルを作成する際より長い文節などの対応関係を取ることを目標とする。

Zha ら [8] によって二次元行列の部分特異値分解に基づくデータ分類手法が提案された。Lardilleux ら [4] はこの手法から階層的な部分文アライメント手法を提案した。

可切性が原言語、目的言語でそれぞれ木構造を作成し対応関係を調べる手法であるのに対し Lardilleux らの手法は二言語間で並列的に構造化しアライメントを行う手法である。本論文ではこの 2 つの手法による翻訳テーブルを比較する。

3　類推翻訳のための翻訳テーブルの生成

3.1　可切性

可切性は文章分割の優先度を表しそれに基づいて文章を分割することである。可切性の値が大きいほどその単語間でのつながりは弱く分割の優先度が高いことを表す。可切性値 sec の計算は次の一般化された式で行う。

例 2. 単語列 [BoS, a, b, c, d, EoS]

$$\mathrm{sec}(\mathrm{bc}) = \frac{p(\mathrm{ab}) \cdot p(\mathrm{bc}) \cdot p(\mathrm{cd})}{p(\mathrm{abc}) \cdot p(\mathrm{bcd})} \quad (1)$$

また、それぞれの p は各バイグラム、トライグラムの出現確率を表し次の式で求められる。

$$p(\mathrm{bc}) = \frac{n(\mathrm{bc}) + \delta}{N + \delta \times V} \quad (2)$$

N はバイグラム、トライグラムの延べ数を表し V はそれぞれの語彙数を表す。本論文では $\delta = 10 \times 10^{-6}$ とした。

分割点前後の単語を含むバイグラム、トライグラムを参照するため文頭、文末での計算では文頭、文末に特別なシンボル（BoS、EoS）を付加した。

[1] http://anymalign.limsi.fr/
[2] http://www.statmt.org/moses/giza/GIZA++.html

計算した値に基づき文章を分割していくことで各接点はそれぞれ可切性の計算値を、終端接点はそれぞれ単語を持つ二分木を構築することができる。次の和文を例に木構造を構築してみる。/は可切性値に基づく分割位置を表す。

例 3. 原言語文: 私 も 知り たい
目的言語文: I also want to know

表 1: 各言語での可切性値

分割点	可切性値
sec(私 も)	0.06
sec(も 知り)	0.26
sec(知り たい)	0.13
sec(I also)	0.11
sec(also want)	0.29
sec(want to)	0.03
sec(to know)	0.05

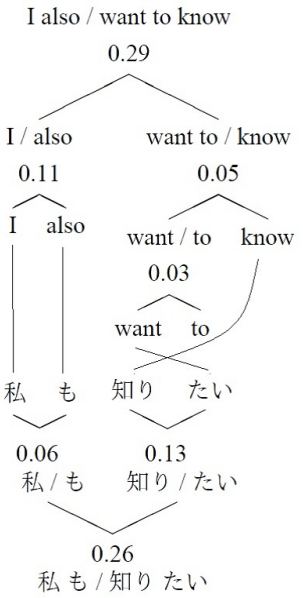

図 1: 可切性による二言語木構造

この様に可切性によって文章を木構造化し意味対応から構造対応が得られる。それぞれの接点における文字列をアライメントとして出力し語彙重み (lexical weight[2])、翻訳確率、出現回数をそれぞれ付加することで翻訳テーブルを作成することができる。

可切性による翻訳テーブルの作成は、

1. 原言語、目的言語でそれぞれ可切性値を計算
2. 可切性値に基づき原言語、目的言語の木構造を作成
3. 単語テーブルを参照し単語の意味対応を調べる
4. 各節で対応する文字列をアライメントとして出力
5. 各アライメントの翻訳確率を計算

といった手順で行われる。参照する単語テーブルは anymalign[3] によって作成した。

3.2 階層的部分文アライメント

作成した翻訳テーブルは翻訳実験によってその品質を評価する。そこで比較対象となるのが Zha(2012) ら [8] の提案した二次元行列の部分特異値分解に基づくデータ分類手法を参考として作成した翻訳テーブルである。

Lardilleux ら [4] はこの手法を利用し二次元行列の縦軸を原言語、横軸を目的言語の単語列と置くことで 2のように各単語列の組み合わせから得られる重みの最大化問題に基づくアライメント手法を提案した。

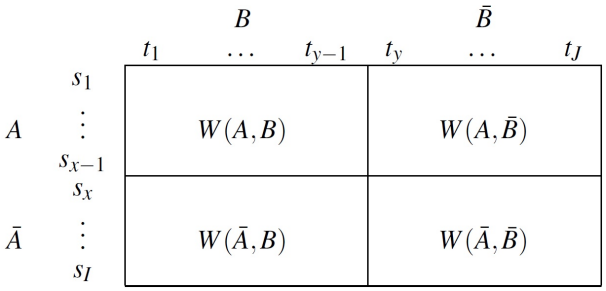

図 2: 文章一対の分割概略図

重みの計算は次の式で行う。

$$W(X,Y) = \sum_{s\in X, t\in Y} w(s,t) \quad (3)$$

$$w(s,t) = p(s|t)\times p(t|s) \quad (4)$$

$$= \frac{\sum_{n=1}^{N}[[(s,t)\in(S_n,T_n)]]k_n}{\sum_{n'=1}^{N}[[s\in S_{n'}]]k_{n'}} \times \frac{\sum_{n=1}^{N}[[(s,t)\in(S_n,T_n)]]k_n}{\sum_{n'=1}^{N}[[s\in T_{n'}]]k_{n'}} \quad (5)$$

$[[x]]$ は真であれば 1、しなければ 0 となる。N は参照翻訳テーブルのエントリー数を表し、$S_n(T_n)$ は参照翻訳テーブルに含まれる原言語または目的言語のエントリーを表す。k_n はテーブル内のペア (S_n,T_n) に関連付けられる回数を表している。

重みが最大となる単語列の組み合わせによって文章を二言語間で並列して階層的に構造化することができる。

この手法では原言語、目的言語で独立した構造を作るのではなく言語間で完全に同期のとれた構造化を行う。

可切性では文章の木構造化を各言語ごとに行う。言語間の構造対応の影響を調べるためこのような二言語間で並列的に文章を構造化するアライメント手法を比較対象とした。

4 翻訳実験

本論文では提案したアライメント手法を評価するため、作成した翻訳テーブルを用例翻訳システムの適用する実験を行う。翻訳結果は BLEU で評価する。

4.1 対象データ

適用する言語対は Europarl parallel corpus release v3[3]のうち標準的な組み合わせである英語-フランス語、言語類似性の低いフィンランド語-フランス語、言語類似性の高いスペイン語-ポルトガル語の三言語対双方向である。それぞれの言語特性およびデータの内訳を次の表に示す。文長は文章を構成する単語数である。

表 2: Europarl Data

訓練セット	347,614 文	
言語	単語数	平均文長 ± 標準偏差
English (en)	9,945,400	29±15
French (fr)	10,959,243	32±17
Finnish (fi)	7,180,028	21±11
Portuguese (pt)	10,302,370	30±16
Spanish (es)	10,482,185	30±17
テストセット	100 文	
言語対	単語数	平均文長 ± 標準偏差
en-fr	2,880	30±10
fr-en	2,638	26±9
fr-fi	1,838	19±7
fi-fr	2,846	29±10
es-pt	2,709	27±9
pt-es	2,747	28±9
チューニング	500 文	

翻訳テーブルは各訓練セットからのみ作成する。

4.2 類推関係に基づいた用例翻訳

類推翻訳システムは Lepage ら [5] によって提案された。まず、類推関係とは A:B::C:D と表現され、A と B の関係は C と D の関係に等しいことを意味する。これにより A、B、C がわかっていれば D も導くことができる。以下に例を示す。

tennis : **I play** tennis :: the piano : D

ここで A と B の違いは太字で示している部分でありこの関係に従うと D は [I play the piano] と導くことができる。

これを利用して翻訳テーブルにない翻訳対象が現れた場合も類似したエントリーを検索して類推関係による解の導出を行うことで翻訳が可能となるのが類推関係に基づく用例翻訳システムである。

本論文で使用した類推翻訳システムは以下の手順で翻訳を行う。

1. 入力文を可切性に基づいて木構造化
2. ボトムアップ方式で小さな部分木から徐々に翻訳
   (a) 対象が翻訳テーブルに存在する場合そのまま翻訳
   (b-1) 対象が翻訳テーブルに存在しない場合類推関係によって導出
   (b-2) 導出した解を翻訳テーブルに追加

このような手順を取ることで徐々に翻訳テーブルの質を向上させながら翻訳することができる。

4.3 実験手順

実験の手順を示す。

1. 訓練セットから anymalign[3] によって参照用の単語テーブルを作成
2. 訓練セット、anymalign による単語テーブルを用いて可切性、階層的アライメント手法それぞれの翻訳テーブルを作成
3. 各翻訳テーブルを使用した類推翻訳システムによってテストセットを翻訳
4. 翻訳結果を BLEU 値によって評価し 2 つの手法による結果を比較

また、GIZA++/MOSES で作成した翻訳テーブルを使用して MOSES による統計翻訳を行った結果を参考値とする。

次にそれぞれの手法を訓練セットに適用した結果得られた翻訳テーブルの大きさを表3に示す。

表 3: 各翻訳テーブルのエントリー数

言語対	可切性	階層的手法
en-fr	1,137,951	2,159,280
fr-en	1,252,256	3,911,212
fr-fi	922,581	1,925,021
fi-fr	765,395	1,448,692
es-pt	1,181,299	3,409,408
pt-es	1,336,049	3,376,451

表 4: 各翻訳テーブルの平均単語数

言語対	可切性	階層的手法
en-fr	14.4±13.9	8.0±9.1
fr-en	14.8±14.6	13.6±14.7
fr-fi	18.6±16.6	18.1±16.6
fi-fr	10.0±9.7	8.6±9.2
es-pt	13.5±14.2	11.0±14.2
pt-es	13.4±13.5	11.0±12.7

表 4は各テーブルに含まれるエントリーの平均単語数を示す。この表から可切性のほうが平均単語数が多い。また、標準偏差から可切性の方ががばらつきが小さく多様な単語数のエントリーを得たことがわかる。

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

4.4 実験結果と考察

各翻訳テーブルを利用して行った翻訳実験の BLEU 評価値を表5に示す。

表 5: 実験結果

言語対	可切性	階層的手法	GIZA++/MOSES
en-fr	**10.2**	9.8	31.30
fr-en	**13.5**	12.1	27.31
fr-fi	**0.4**	0.3	11.85
fi-fr	**1.0**	0.2	14.36
es-pt	23.7	**23.9**	32.53
pt-es	**20.9**	17.6	34.94

スペイン語からポルトガル語への翻訳を除くと可切性によって作成した翻訳テーブルを使用した方が良好な BLEU 値を得られている。

この結果から類推翻訳向けの翻訳テーブルでは並列的に構造化しアライメントを導くよりも言語ごとの構造を維持した上でアライメントを行うほうが適切であるといえる。また、表 4も考慮すると、より多様な単語数のエントリーを持つことは翻訳システムの質を向上させることがわかる。

ただし、今回我々が使用した類推翻訳システムは入力文の構造化を可切性によって行っているため階層的アライメント手法はシステムとの適合性において不利であったことは確かである。階層的アライメント手法は二言語間で並列的に構造化するため入力が単言語となる翻訳システムではその性能を最大限に発揮することは難しい。

しかしながらどちらの手法も統計的機械翻訳の結果には遠く及ばず、さらなる手法の改善や新たな手法の提案が必要である。

5 おわりに

ここまで用例翻訳システムを対象とした翻訳テーブルの作成手法として可切性に基づくアライメントについてを述べてきた。翻訳実験では比較対象として言語間で並列化した構造化を行う事のできる階層的アライメント手法を利用した。これによって言語の構造化を各言語で独立して行う場合と言語の組み合わせから行う場合のどちらが類推翻訳システムに適しているかを検討した。

本論文で行った翻訳実験からは各言語ごとに構造化を行い文部分的な対応関係を探索した可切性による翻訳テーブルの方が類推翻訳システムに適している、という結果が得られた。ただし、この優勢は我々の翻訳システムが同じく可切性によって文章を構造化するためであるとも考えられる。

翻訳テーブルの品質向上だけでなく翻訳システムの改良も含めた検討が必要である。

謝辞

本研究は JSPS 科研費 基盤 C 23500187 の助成を受けたものである。また、本稿は早稲田大学特定課題研究助成費（課題番号 2013A-6336）による研究成果の一部である。

参考文献

[1] Christophe Chenon. *Vers une meilleure utilisabilité des mémoires de traduction, fondée sur un alignement sous-phrastique*(**翻訳メモリのよりよい活用に向けた文部分的アライメント**)[**仏題**]. PhD thesis, Docteur de l'universite Joseph Fourier, 2005.

[2] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[3] Adrien Lardilleux, Yves Lepage, et al. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing*, pages 214–218, 2009.

[4] Adrien Lardilleux, François Yvon, and Yves Lepage. Alignement sous-phrastique hirarchique avec anymalign (hierarchical sub-sentential alignment with anymalign) [in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 113–126, Grenoble, France, June 2012. ATALA/AFCP. URL `http://www.aclweb.org/anthology/F12-2009`.

[5] Yves Lepage and Etienne Denoual. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282, December 2005. ISSN 0922-6567.

[6] Franz Josef Och and Hermann Ney. Improved statistical alignment models. pages 440–447, Hongkong, China, October 2000.

[7] Kota Takeya. Analogy-based transratoin:use of marker-based chunking and secability. Master's thesis, Graduate School of Infomation, Production and Systems Waseda University, 2012.

[8] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 25–32, New York, NY, USA, 2001. ACM. ISBN 1-58113-436-3.