

# 類推関係に基づいた用例翻訳のための翻訳テーブルの生成と評価

木村 竜矢<sup>†1</sup>      松岡 仁<sup>†2</sup>      西川 裕介<sup>†3</sup>      ルパージュ・イヴ<sup>†4</sup>

早稲田大学 理工学術院 大学院情報生産システム研究科<sup>†</sup>

## 1 はじめに

類推関係に基づいた用例翻訳手法はルパージュら [1] によって提案された。この手法では類推関係に基づいて翻訳を行うため類推関係が成り立ちにくい入力文に対しての翻訳は困難である。ここで使用される類推関係とは A と B の関係が C と D の関係に等しいという意味である。類推関係の表記として  $A : B :: C : D$  と表現する。例えば、 $A = \text{“I’ve always had a dog since I was small.”}$   $B = \text{“I prefer dogs to cats because I’ve always had a dog since I was small.”}$   $C = \text{“I’ve always drunk tea since I was small.”}$   $D = \text{“I prefer tea to coffee because I’ve always drunk tea since I was small.”}$  の文があったとする。もし、A と B と C の翻訳がわかっていたら D の翻訳を類推関係より導出可能である。しかし、A と B と C の入力に翻訳テーブルから検索するため、前例のように長い文は既存の翻訳テーブル (GIZA++ [3], Anymalign [4] など) には存在しない。そのため、本論文では類推関係に基づいた用例翻訳システムの為の新たな翻訳テーブルの生成手法について述べる。

## 2 可切性と翻訳テーブルの生成

### 2.1 可切性

可切性とは文 (テキスト) において区切りやすさを意味する。可切性は機械翻訳の為にシュノンら [2] によって提案された。可切性の値が高ければ高いほど、その点で文が区切れることを示している。連続する 2 つの単語  $e_i$  と  $e_{i+1}$  の間の可切性の値は以下の式 1 により計算できる。しかし、訓練データ中に  $e_i e_{i+1}$  が存在しなければ  $C(e_{i-1}e_i)$  の出現数が 0 となり、確率値を計算することは不可能である (ゼロ頻度問題)。本論文では上記の問題を解決するためラプラスのスムージ

ング手法を用いる (式 2)。

$$\text{sec}(e_i e_{i+1}) = \frac{p(e_{i-1}e_i) \cdot p(e_i e_{i+1}) \cdot p(e_{i+1}e_{i+2})}{p(e_{i-1}e_i e_{i+1}) \cdot p(e_i e_{i+1} e_{i+2})} \quad (1)$$

$$p(e_{i-1}e_i) = \frac{C(e_{i-1}e_i) + 1}{N + V} \quad (2)$$

ここで、 $C(e_{i-1}e_i)$  は訓練データでの  $e_{i-1}e_i$  の出現数を示す。 $N$  は訓練データでの全ての N-gram の総数、 $V$  は訓練データでの全ての N-gram の異なり数を表す。

### 2.2 翻訳テーブルの生成

まず、単語間のアライメントを Anymalign[4] を用いて抽出する (図 1 破線四角)。Anymalign は語彙抽出において従来の代表ツールである GIZA++[3] より性能が優れている事が報告されている [5]。その後、元言語と目的言語をそれぞれ独立で可切性の計算を行い木構造を生成する。木構造でアライメントを取る類似手法 ITG (Inversion Transduction Grammar) [6] に対して、本手法では得られた木構造から図 1 のように単語間のアライメント情報を付加することで元言語と目的言語の翻訳関係を得ることができる (図 1 の四角で示した部位)。ここで得られる翻訳関係は “le fruit” と “the fruit”、“mange le fruit” と “eats the fruit”、“mange le fruit” と “the fruit”、“le fruit” と “eats the fruit” の 4 つの翻訳関係である。この手法を長い文で適用することによって、長い単位の翻訳関係を抽出することが可能である。

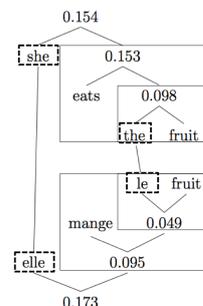


図 1: 翻訳関係の抽出

Evaluation of translation tables produced using secability in an example-based machine translation system by analogy

<sup>†1</sup> Tatsuya Kimura      <sup>†2</sup> Jin Matsuoka

<sup>†3</sup> Yusuke Nishikawa      <sup>†4</sup> Yves Lepage

<sup>†</sup> Graduate School of Information, Production and Systems, Waseda University, Japan

表 1: 類推関係に基づいた用例翻訳システムによる翻訳テーブルの評価

	fr-en		en-fr		fi-fr		fr-fi		pt-es		es-pt	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
可切性	<b>13.5</b>	<b>0.73</b>	<b>10.2</b>	<b>0.71</b>	0.4	<b>0.88</b>	<b>1.0</b>	<b>1.24</b>	<b>23.7</b>	<b>0.58</b>	<b>20.9</b>	<b>0.61</b>
Anymalign	12.2	0.78	8.9	0.74	<b>1.1</b>	1.23	<b>1.0</b>	<b>1.24</b>	22.1	0.62	20.5	0.63
GIZA++/ Moses	0.4	0.96	0.8	0.95	0.1	1.00	0.1	1.25	5.7	0.87	7.6	0.83

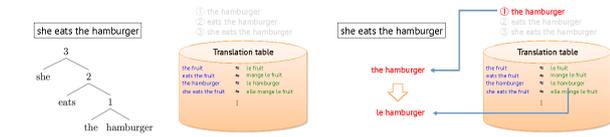


図 2: 可切性による木構造の生成

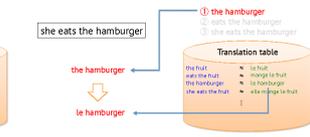


図 3: 翻訳テーブルを用いた翻訳

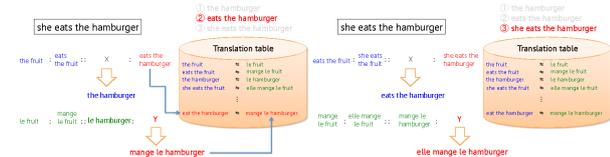


図 4: 新たに生成された言語資源を追加

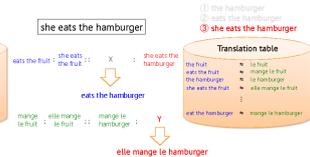


図 5: 新たに生成された翻訳テーブルを用いての翻訳

### 3 類推関係に基づいた用例翻訳

本論文では可切性で得られた翻訳テーブルを類推関係に基づいた用例翻訳システムに適用する。以下に例として、“she eats the hamburger”を英語からフランス語に翻訳する過程について説明する。まず、図2のように可切性を計算し文章を木構造化する。次に、その木構造に基づいて文章を図2の①から③のように分割する。その後、小さな部分木から類推関係に基づいて翻訳を行う。ここでは、“the hamburger”から翻訳を始める。翻訳したいものが翻訳テーブルに存在する場合、そのまま翻訳結果を出力する(図3)。翻訳したいものが翻訳テーブルに存在しない場合、類推関係に基づいて翻訳を行い、結果は新たに翻訳テーブルに追加する(図4)。最後に“she eats the hamburger”を類推関係に基づいて翻訳する(図5)。以上に示したように、小さい部分木から翻訳を行うことで翻訳テーブルの質が翻訳を行うにつれ高くなり、翻訳できる可能性を向上させる事ができる。

### 4 実験と評価

本論文の実験では、Europarl コーパス [7] を使用した。訓練データは 347,614 文で、テストデータは 100 文である。なお、チューニングは行わない。言語ペアとしては、ヨーロッパの主要な 11 言語において、通常最も BLEU スコア [8] が低いフランス語とフィンランド語と、通常最も BLEU スコアが高いスペイン語とポルトガル語、それに加え、英語とフランス語に関して

実験を行った。比較のため、合わせて3つの異なる翻訳テーブルを類推関係に基づいた用例翻訳システムを使用して翻訳実験を行った。1つめは提案手法である可切性を用いた翻訳テーブル、2つめは Anymalign による翻訳テーブル、最後は GIZA++/Moses による翻訳テーブルである。表2は翻訳テーブルのエントリー数とその長さの関係を示す。表1では3つの異なる翻訳テーブルを類推関係に基づいた用例翻訳システムを使用して翻訳した結果を BLEU[8] と TER[9] で評価した結果である。

表 2: 翻訳テーブルのエントリー

	fr-en	fr-fi	pt-es
可切性: エントリー数	1,255,774	1,045,670	1,337,194
長さ: 加算平均 ± 標準偏差	10.04±14.96	11.40±17.40	10.47±15.56
Anymalign: エントリー数	1,063,042	1,033,599	997,805
長さ: 加算平均 ± 標準偏差	1.26±0.83	1.27±0.91	1.28±0.86
GIZA++: エントリー数	1,329,320	597,235	1,681,573
長さ: 加算平均 ± 標準偏差	3.03±1.27	3.03±1.47	3.23±1.37

### 5 考察と結論

本論文では類推関係に基づいた用例翻訳において翻訳の可能性を高めるための新たな翻訳テーブルの生成手法について提案した。本手法では平均 10 単語以上の長さを持つ翻訳テーブルのエントリーの生成に成功した(表2)。また、類推関係に基づいた用例翻訳実験では BLEU スコアにおいても、TER スコアにおいても、その有効性を実証する事が出来た(表1)。

### 参考文献

- Y. Lepage and E. Denoual. Puresst ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3):251-282, 2005.
- C. Chenon. Vers une meilleure utilisabilité des mémoires de traduction, fondée sur un alignement sous-phrastique *PhD thesis*, Université Joseph Fourier-Grenoble 1, 2006.
- F.J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19-51, 2003.
- A. Lardilleux and Y. Lepage. Sampling-based multilingual alignment. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214-218, 2009.
- A. Lardilleux, J. Gosme, and Y. Lepage. Bilingual Lexicon Induction: Effortless Evaluation of Word Alignment Tools and Production of Resources for Improbable Language Pairs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, pages 252-256, Valletta, Malta, 2010.
- D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377-403, 1997.
- P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79-86, Phuket, Thailand, 2005.
- G. Papineni, S. Roukos, T. Ward, and W.J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL2002)*, pages 311-318, 2002.
- M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the Xth annual meeting of the Association for Machine Translation in the Americas (AMTA2006)*, pages 223-231, 2006.