

用例ベース翻訳の確率的モデル化

荒牧 英治[†] 黒橋 禎夫^{††}
柏岡 秀紀^{†††,††††} 加藤 直人^{††††}

用例ベース翻訳は、これまで、経験則にもとづく指標／基準により用例を選択してきた。しかし、経験則に頼った場合、その修正を行うのが困難であり、また、アルゴリズムが不透明になる恐れがある。そこで、本研究では用例ベース翻訳を定式化するための確率モデルを提案する。提案するモデルは、翻訳確率の最も高い用例の組み合わせを探索することで、翻訳文を生成する。さらに、本モデルは用例と入力文のコンテキストの類似度を自然に翻訳確率に取り込む拡張も可能である。実験の結果、本モデルを用いたシステムは、従来の経験則によるシステムの精度を僅かに上回り、用例ベース翻訳の透明性の高いモデル化を実現することに成功した。

キーワード： 用例ベース翻訳，機械翻訳，確率モデル，コンテキストの類似，依存構造

Probabilistic Formalization for Example-based Machine Translation

ELJI ARAMAKI[†], SADA O KUROHASHI^{††}, HIDEKI KASHIOKA^{†††,††††}
and NAOTO KATO^{††††}

Example-based machine translation (EBMT) systems, so far, rely on heuristic measures in retrieving translation examples. Such a heuristic measure costs time to adjust, and might make its algorithm unclear. This paper presents a probabilistic model for EBMT. Under the proposed model, the system searches the translation example combination which has the highest probability. The proposed model clearly formalizes EBMT process. In addition, the model can naturally incorporate the context similarity of translation examples. The experimental results demonstrate that the proposed model has a slightly better translation quality than state-of-the-art EBMT systems.

KeyWords: *Example-based Machine Translation, Machine Translation, Probabilistic Model, Context Similarity, Dependency Structure*

1 はじめに

近年、統計ベース翻訳 (Brown, Pietra, cent J. Della Pietra, and Mercer 1993) や用例ベース翻訳 (Nagao 1984) など大量のテキストを用いた翻訳手法 (コーパスベース翻訳) が注目され

[†] 東京大学附属病院企画情報運営部, Department of Planning, Information and Management, University of Tokyo Hospital

^{††} 京都大学大学院情報学研究所, Graduate School of Informatics, Kyoto University

^{†††} 独立行政法人情報通信研究機構, National Institute of Information and Communications Technology

^{††††} ATR 音声言語コミュニケーション研究所, ATR Spoken Language Translation Research Laboratories

^{†††††} NHK 放送技術研究所, Science and Technical Research Laboratories of NHK

ている。我々は、用例ベース翻訳に焦点を当て研究を行っている。

用例ベース翻訳の基本的なアイデアは、入力文の各部分に対して類似している用例を選択し、それらを組み合わせて翻訳を行うことである。ここでいう類似とは、通常、入力文とできるかぎり大きく一致していればいるほど（一致する単語数または文節数が多いほど）よいと考えられてきた。これは、用例が大きくなればなるほど、より大きなコンテキストを扱うことになり、正確な訳につながるからである。そのため、これまでの用例ベース翻訳は、大きな用例を優先する指標／基準を用いて用例を選択してきた。

一方、統計ベース翻訳は、翻訳確率を緻密に計算するため、基本的には、翻訳用例を小さな語／句単位に分解して学習を行う。もちろん、最近の統計ベース翻訳では、より大きな単位を取り扱う試みも行われている。例えば、Och(Och, Tillmann, and Ney 1999)等は、アライメントテンプレートという単位を導入し、語列をまとめて学習した。また、他にも多くの統計翻訳研究が語よりも大きな単位を学習に取り込む試みを行っている(Koehn, Och, and Marcu 2003; Watanabe, Sumita, and Okuno 2003)。しかし、入力文とできる限り大きく一致した用例を用いる用例ベース翻訳と比べると、あらかじめ翻訳単位を設定する統計ベース翻訳の扱う単位は依然として小さいと言える。

簡単に言うと、統計ベース翻訳と用例ベース翻訳は、以下の2点で異なる。

- (1) 用例ベース翻訳は、用例のサイズ（一致する単語数または文節数）を重視している。統計ベース翻訳は用例の頻度を重視している。
- (2) 用例ベース翻訳は、経験則による指標／基準にもとづいて動作している。統計ベース翻訳は確率的に定式化されている。

本研究では、用例ベース翻訳の問題は、(2) 経験則による指標／基準を用いている点だけと考える。経験則による指標／基準は、調整や修正が困難であり、また、アルゴリズムが不透明になる恐れがある。

そこで、本研究では、用例ベース翻訳を定式化するために、用例ベース翻訳のための確率モデルを提案する。提案する翻訳モデルは、統計ベースのそれとは異なり、語や句単位の小さな単位から、文全体まで、あらゆるサイズをカバーした翻訳確率を計算する。この枠組みの上では、大きなサイズの用例は安定した翻訳先を伴うため、高い翻訳確率を持つと考えられる。したがって、翻訳確率が高い用例を選ぶことで、自然と用例のサイズを考慮した用例の選択が可能となる。

また、提案する翻訳確率は容易に拡張可能であり、用例と入力文のコンテキストの類似度を確率モデルに取り込むことも可能である。

実験の結果、提案手法は、従来の経験則に基づいた翻訳システムよりも僅かに高い精度を得て、用例ベース翻訳の透明性の高いモデル化を実現することに成功したので報告する。

提案手法は言語ペアを特定しないが、本稿は日英翻訳方向で説明し、実験を行った。

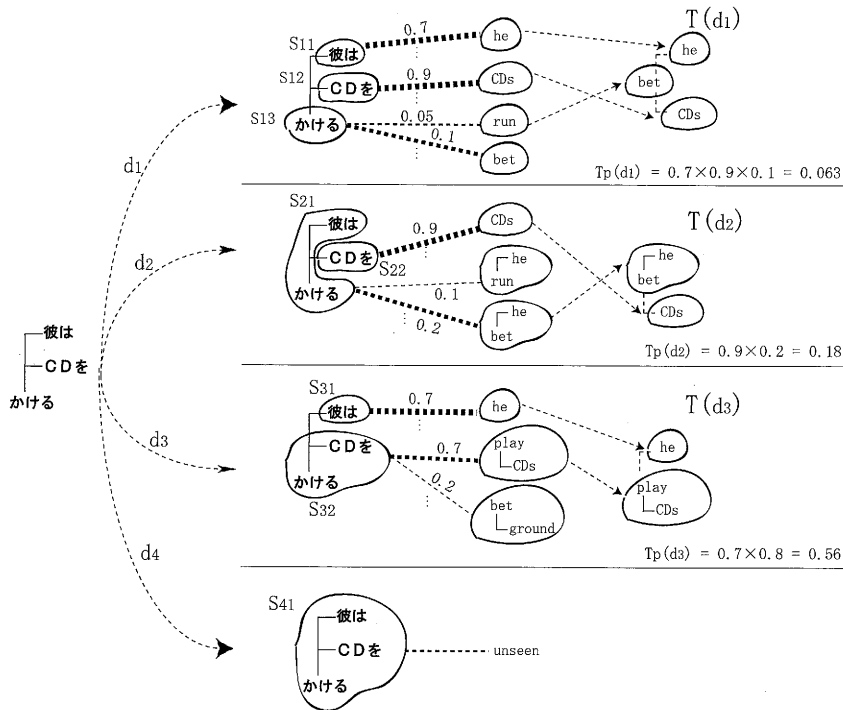


図 1 翻訳のながれ

* 本稿の図では、依存構造木の親を左に、子を右に描く。また、本研究では、ノードの単位は日本語は文節、英語は base-NP、または base-VP とする。

本稿の構成は、以下のとおりである。2章では、提案するモデルの基本的アイデアについて説明する。3章では、アルゴリズムについて述べる。4章では、実験について報告する。5章では、関連研究を紹介し、6章に結論を述べる。

2 提案手法

用例ベース翻訳の基本的な原則はできるだけ大きなサイズの用例を用いて翻訳文を生成することである。これを確率的に定式化するためには、大きな用例を用いた翻訳結果が大きな翻訳確率を持たなくてはならない。本章では、これを実現するための基本アイデアを述べる。

まず、提案手法は入力文を可能なかぎりの部分木の組合せに分解する：

$$D = \{d_1, \dots, d_N\}. \tag{1}$$

ここで、 d_i は入力文の分解のパターン、 D は d_i の集合である。

次に、 d_i は入力文を M_i 個の部分木に分解しているとする：

$$d_i = \{s_{i1}, s_{i2}, \dots, s_{iM_i}\}, \quad (2)$$

ここで、 s_{ij} は入力文の部分木である。

例えば、図 1 左の入力文の場合、 d_1, \dots, d_4 の 4 通りの部分木の組合せで表現できる。この例では、 d_1 は入力文を 3 つの部分木 s_{11}, s_{12}, s_{13} に分解している。また、 d_2, d_3 は、入力文を 2 つの部分木に分解している。また、 d_4 のように、文そのものも分解パターンとして取り扱う。

次に、各部分木 s_{ij} それぞれについて、もっとも翻訳確率 $P(t_{ij} | s_{ij})$ (この確率の計算方法は次節にて述べる) の高い用例を選び、それらの積を翻訳文の翻訳確率 $T_P(d_i)$ とする：

$$T_P(d_i) = \prod_{s_{ij} \in d_i} P(t_{ij} | s_{ij}). \quad (3)$$

ここで、 t_{i1}, \dots, t_{iM_i} を d_i の翻訳とみなし、 $T(d_i)$ と表記する。

最後に、もっとも高い翻訳確率を持つ分解パターン (d_m) を以下の式によって探索し、最終的な翻訳を $T(d_m)$ とする：

$$d_m = \arg \max_{d_i \in D} T_P(d_i). \quad (4)$$

簡単に言うと、提案手法は、入力文のある単位をどう翻訳するかと、どういう単位で翻訳するかという 2 つ問題を解いている。前者は、最も確率の高い用例を選択することで解決しており、基本的な統計的翻訳と同様の考え方である。後者は、入力文の分解パターンを選択することで解決している。

ここで重要なことは、本モデルの枠組みでは、大きな用例を用いた翻訳文が優先されることである。この理由は、大きな用例は安定した翻訳先を持つ傾向にあるため、高い翻訳確率を持ち、当然、その積である翻訳文の確率 $T_P(d_i)$ も自然と高くなるからである。

例えば、日本語“かける”は、翻訳の際には大きな曖昧性があり、“bet”、“run”や“play”など様々な英語表現が考えられる。

ここで、もし、 $T(d_1)$ のように、入力文を小さな部分木に分解した場合は、適切な訳である $P(\text{play} | \text{かける})$ の翻訳確率は低く、適切な翻訳は行われない。

一方、 $T(d_2)$ では、より大きな表現“CD をかける”を用いた用例を探索している。この用例の英語表現としては、ほとんどが“play”となり、用例の翻訳確率は高くなる。その結果、用例群の翻訳確率の積である $P(d_2)$ も高くなり、この結果が翻訳として採用される。

また、図 1 の $T(d_4)$ のように、大きすぎる単位で検索した場合は、コーパス中に存在せず、確率が定義されない場合がある。

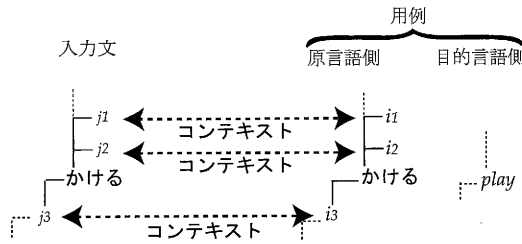


図 2 コンテキストの定義

表 1 “かける”を含んだ用例とそのコンテキスト (コンテキストは括弧で示されている)

用例原言語側	用例目的言語側	context.sim
(テープを) かける	play	0.8
(CDを) かける	play	0.8
(MDを) かける	put	0.8
(目覚ましを) かける	set	0.6
:	:	:
(お金を) かける	bet	0.4
(1万円を) かける	bet	0.4
(名誉を) かける	bet	0.3

* 実際には用例は木構造の形で扱われているが, 表記を簡潔にするため, この図では用例の構造は記していない。

2.1 パラメータの推定

本節では, 用例の翻訳確率の推定方法を述べる. まず, 英語部分木 t と日本語部分木 s からなる用例があるとす. この翻訳確率 $P(t | s)$ は, アライメントされたコーパス中での対応 (t, s) の出現頻度を直接数えて求める:

$$P(t | s) = \frac{\text{count}(t, s)}{\text{count}(*, s)}, \tag{5}$$

ここで, $\text{count}(t, s)$ は, アライメントされたコーパスにおける対応 (t, s) の出現頻度, $\text{count}(*, s)$ は日本語部分木 (s) の出現頻度である¹.

ただし, この頻度の計算にあたっては, 次節に述べるコンテキストの情報を利用した拡張が可能である.

¹ 後述する実験では, データスパースネスの問題に対処するため, s と t は内容語に汎化して集計を行った.

2.2 入力文と用例のコンテキストの類似度を取り込んだ確率モデル

用例の選択にあたって重要な手がかりは入力文と用例の一致するサイズであり、それは、2.1節で提案された翻訳確率の枠組みで実現されている。しかし、用例のサイズに加えて、入力文と用例のコンテキストの類似も重要な手がかりである。提案するモデルは、このようなコンテキストの類似を取り込む拡張を自然に行うことができる。

まず、提案手法を説明する前に、用例と入力文のコンテキストを定義する。図2に示されるように、用例の原言語側が $i_{1..3}$ という3つの句と接続しているとする。これらとそれと対応する入力文の $j_{1..3}$ をコンテキストと考える。

そして、用例と入力文のコンテキストの類似度を次の式で定める：

$$\text{context_sim}(s) = \sum_{i \in N} \text{sim}(i, j), \quad (6)$$

ここで、 i は用例 A の日本語側で翻訳に使う部分の周辺の句、 j は i と対応する入力文の句、 N は i の集合、 $\text{sim}(i, j)$ はシソーラス (池原悟, 宮崎正弘, 白井論, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 1997) を用いて計算する i と j の類似度 ($\max=1$) であり、以下の式で定義される：

$$\text{sim}(i, j) = \frac{2d_c}{d_i + d_j}, \quad (7)$$

ここで、 d_i と d_j は、それぞれ i と j のシソーラス上での深さ、 d_c は、 d_i と d_j の共通するパスの深さである。

提案手法のポイントは、高い類似度を持つ用例は、同じく高い類似度を持つ用例のみを用いて翻訳確率を計算する点である。すなわち、式5によって、ある用例の翻訳確率を計算する際には、 $\text{context_sim}(s)$ 以上の類似度を持つ用例だけを集計して翻訳確率を計算し、 $\text{context_sim}(s)$ 未満の類似度の用例は、用例の翻訳確率の計算には用いない。この操作を用例のフィルタリングと呼ぶ。

このフィルタリングの操作は、用例のサイズごとに翻訳確率を計算する手法を、類似度までに拡大したものであり、自然な拡張であるといえる。この拡張の結果、高い context_sim を持つ用例は、それよりも低い context_sim を持つ用例の影響を受けず、多くの場合、高い翻訳確率を持つことになる。

例えば、“レコードをかける”がコーパスに存在しない（しかし、“レコード”と“かける”それぞれ単独では出現している）場合に、表1の用例群を用いて翻訳することを考える。前節までの手法では、このように、大きなサイズで一致するものがない場合、“かける”単独で翻訳確率を計算することになり、“bet”など不適切な訳語が選ばれる可能性がある。

本節の提案手法では、用例“CDをかける”と入力文“レコードをかける”の context_sim が0.8であるとすると、同じく0.8以上の context_sim を持つ用例だけを用いて翻訳確率を計算す

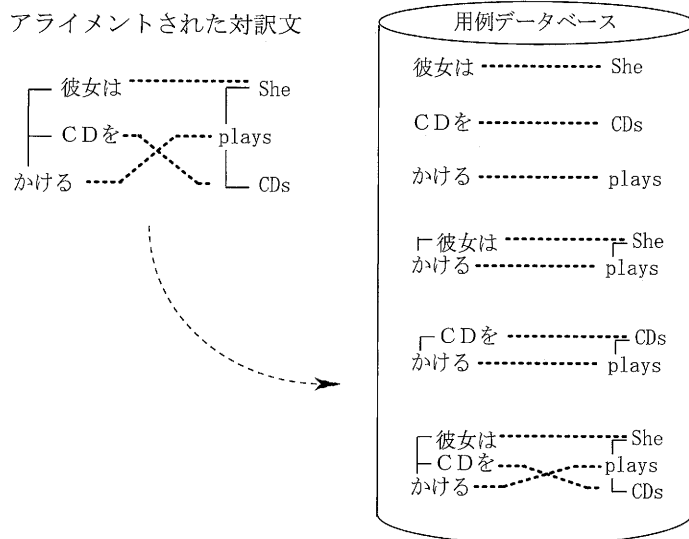


図 3 用例データベースの構築

る. この場合, 用例の数は3つだけとなるとが, その英語表現は安定しており, $P(play | かける) = \frac{2}{3}$, $P(put | かける) = \frac{1}{3}$ となる. このように類似したコンテキストを持つ用例の翻訳確率は自然と高くなる傾向をもつ.

また, この枠組では, 類似度をもっとも高い用例が一つしかない場合, その翻訳確率は最大の1となる. これは, より大きな用例が利用可能であった場合に, その大きな用例よりも, 類似している小さな用例を優先しているかのように見える. この問題は次のように解決されている.

まず, 提案手法は用例を構築する際に, 大きな用例を分解した一部分も独立した用例として扱い, データベースに蓄える (3.1節ステップ3). よって, ある大きな用例が利用できる状態で, それよりも小さい, もっとも類似した用例が一つしかない場合における, その一つしかない用例とは, 大きな用例の一部分から作られた用例となる. というのは, 大きな用例が利用可能であるならば, その分解から得られた用例は, その周辺が入力文と同一であり, 最大の類似度とるためである. よって, この場合, 大きな用例ともっとも類似している用例のどちらを採用すると考えても, 作られる翻訳は同じとなる.

3 翻訳システムの構成

提案するシステムは, 次の2つのモジュールから構成される:

- (1) アライメント・モジュール: コーパスから用例を構築するモジュール,

(2) 翻訳モジュール：翻訳を行うモジュール.

3.1 アライメント・モジュール

ステップ1：対訳文の依存構造への変換

まず、対訳文を日本語パーサ KNP (Kurohashi and Nagao 1994) と英語パーサ nl-parser(Charniak 2000) によって統語解析する.

日本語の句の単位は、KNP の出力する文節とし、KNP の出力する依存構造をそのまま以降の処理に用いる.

英語パーサは句構造を出力するので、句構造中の主辞を決定して、出力結果を依存構造に変換する. この際、主辞の決定には人手で作成した規則を用い、句の単位は base-NP, base-VP とした.

ステップ2：アライメント

次に、翻訳辞書を用いたアライメントを行い、両言語の句の対応関係を得る. これには、(Aramaki, Kurohashi, Sato, and Watanabe 2001) の手法をそのまま用いた. この手法は、辞書を使用するが、後述する実験では次の辞書を用いた：EDR 電子化辞書², EDICT³, 英辞郎⁴. これらの辞書はのべ二百万項目を持つ.

このステップの結果、システムは、図3左のようなアライメントされた対訳文を得る.

ステップ3：用例データベースの構築

最後に、アライメントされた対訳文(図3左)から、用例データベースを構築する. この際、システムは、あらゆる対応の組み合わせを生成し、その周辺の句(これはコンテキストの類似度を計算する際に用いる)とともにデータベースに登録する(図3右).

3.2 翻訳モジュール

ステップ1：入力文の解析

まず、入力文を日本語パーサ KNP(Kurohashi and Nagao 1994) を用いて統語解析する.

ステップ2：用例の選択

入力文のあらゆる可能な部分木の組み合わせに分解し(前章の図1の左), それらの部分木それぞれについて、用例データベース中を検索し、前章の手法にて、その翻訳確率を計算する. そして、最も翻訳確率の高くなる用例の組み合わせを採用する.

2 http://www2.nict.go.jp/kk/e416/EDR/J_index.html

3 <http://www.csse.monash.edu.au/jwb/j-edict.html>

4 <http://www.eijiro.jp/>

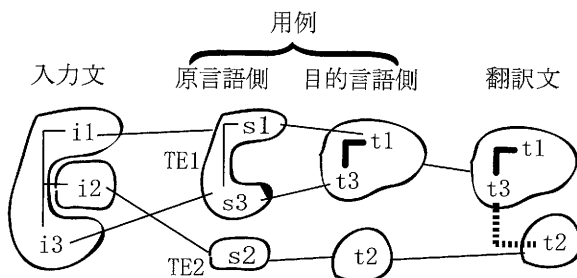


図 4 出力文の生成

ステップ 3: 翻訳文の生成

前ステップで採用された用例群を結合し, 出力文の依存構造にまとめ上げる. この操作は次の 2 つの規則によって行われる.

- (1) 用例内部の依存構造は, 出力文にそのまま用いる. 例えば, 2 つの翻訳用例 (TE1, TE2) を結合して翻訳文を生成する場合を考える (図 4). ここで, TE1 に含まれる 2 つの句 (t_1, t_3) の依存関係 (太線で描かれている) は, そのまま翻訳文に用いられる.
- (2) 用例間の依存関係は, その用例が対応する入力文句の依存構造と同じ親子関係とする. 例えば, 図 4 の TE1 と TE2 の間の依存関係について考える. TE2 は入力文の i_2 と対応しており, i_2 は i_3 と親子関係にある. よって, 翻訳文側でも, i_2 と対応する t_2 は, i_3 と対応する t_3 と親子関係にあるものと考え, 点線で描かれている依存関係を得る.

最後に, 出力文の依存構造の語順を決定する. これは, 先の依存構造の場合と同様に, 用例の内部の語順は保存し, 用例のつなぎ目の語順は, 単語 n -gram 言語モデル⁵にて優先される語順を採用する.

⁵ 単語 n -gram の学習は, 後述する実験のトレーニングセット 2 万文を用いて $n = 3$ にて行った.

4 実験

4.1 実験設定

提案手法の妥当性を検証するため、用例ベース翻訳システム (Aramaki and Kurohashi 2004) の用例選択部分を提案手法に置き換えて実験を行った。

コーパスは IWSLT04 (Akiba, Federico, Kando, Nakaiwa, Paul, and Tsujii 2004) にて配布されたコーパス (トレーニングセットとテストセット) を用いた。トレーニングセットは旅行対話ドメインの 2 万の日英対訳文からなる。テストセットは日本語文 (500 文) とそれらの 16 通りの英語翻訳 (500 × 16 文) からなる。

精度を比較した翻訳システムは次のとおりである。

- (1) PROPOSED: 提案手法。
- (2) WITHOUT_SIM: シソーラスを用いない (コンテキストの類似を扱わない) 提案手法。
- (3) BASIC: 経験則によるメジャーにより用例を選択するシステム (Aramaki and Kurohashi 2004)。このシステムは、IWSLT04 (Akiba et al. 2004) に参加し、高い翻訳精度を示した。また、このシステムは、PROPOSED と同じアライメント結果を用いている。
- (4) BASELINE: 用例ベース翻訳のベースライン。このシステムは、最も編集距離が近い用例を探索し、その用例の英語文をそのまま出力する。
- (5) C1, C2: ルールベース方式の商用翻訳システム。

4.2 評価

評価は、表 3 の自動評価尺度を用い、IWSLT04 (Akiba et al. 2004) と同様の以下の条件で行った。

- (1) 大文字/小文字の差異の無視。
- (2) 記号/句読点 (‐, ?) の無視。
- (3) 数字はスペルアウトする (20,000 → Twenty Thousand)。

4.3 結果

各手法の精度を表 4 に示す。表 4 に示されるように、提案手法 PROPOSED は、経験則による BASIC と比べて僅かに高い精度を持ち、提案する確率による選択が妥当であることを示している。

また、両者の精度は、商用システム (C1, C2) や用例ベース翻訳のベースライン (BASELINE) と比べてはるかに高く、現実的な精度上での比較であることが分かる。

次に, コンテキストの類似度を考慮した効果について述べる. これは, 提案手法 (PROPOSED) とシソーラスを用いない手法 (WITHOUT_SIM) の精度を比較することで調査できる (表 4). 実験結果は, NIST においては PROPOSED が高く, BLEU においては WITHOUT_SIM が僅かに高かった. NIST は訳語選択に鋭敏な自動評価手法であるが, この NIST で, PROPOSED が高い値を持つのは, コンテキストの類似度の導入が訳語選択に貢献しているためだと推測される.

結果をより具体的に比較するため, 両手法で翻訳結果が異なった例の一部を表 2 に示す.

表最上部は「ありませんか」という訳し分けが必要な表現の例である. この表現は, 旅行ドメインでは, 通常は “do you have” と訳すことが多いのだが, 「(施設/場所が) ありませんか」という場合にはこの訳は不適切となる. このような場合でも, PROPOSED は, 「ありませんか」を “are there” 用いて正しく訳せている. 一方, コンテキストの類似度を考慮しない WITHOUT_SIM は, “do you have” と不適切な翻訳を行ってしまう.

このように一部の翻訳で, PROPOSED がより適切な翻訳を行ったが, 表中段の “notify” と “contact” の差異など, 両手法のどちらの訳語が適切か判断が困難な例も数多く確認された. この実験は, ドメインを旅行対話に絞った翻訳実験であり, 訳し分けを必要とする入力文が, ドメインを絞った時点で減っているのが, 一因であると考えられる.

最後に, 表 2 下部のように PROPOSED の方が誤った訳を出力する場合も観察された. これは, PROPOSED の選んだ用例にアライメント誤りが含まれていることが原因であった. 一方, WITHOUT_SIM は, 同じ用例のサイズであれば, 対応する頻度が高い用例を選ぶ. 複数の用例がみな同じアライメントの誤り方をするわけではないので, WITHOUT_SIM は必然的にアライメント誤りの影響を受けにくいと考えられる. この違いが原因で, 一部の自動評価指標においては, WITHOUT_SIM の精度の方が PROPOSED よりも高くなったと推測される.

このように, 今回の実験では, PROPOSED がすべての面で優位であることを示すことはできなかった. しかし, (1) アライメント誤りによる精度の低下は, モデルの定式化の妥当性とは別個の問題である点, また, (2) 実験は, ドメインを絞った翻訳実験であり, コンテキストを考慮する必要性が少ない点, これらの 2 点を考慮すると, コンテキストの類似度の定式化の妥当性は, 実験によって確かめられたと考えられる. また, アライメント誤りに対する頑健性をどのように PROPOSED に持たせるかは, 今後の課題としたい.

4.4 誤り分析

PROPOSED のより具体的な分析のため, PROPOSED の翻訳結果から, 100 翻訳文を無作為抽出し, それらを人手でチェックした.

この結果, 49 の翻訳文が正解であり, 51 の翻訳文が不正解であると判定された. 不正解であった原因を人手で分類した結果を表 5 に示す.

表 5 に見られるように, DATA-SPARSENESS がもっとも顕著な問題である. このことか

表 2 proposed と without_sim の違い

入力文	この近くでいいレストランはありませんか
PROPOSED	are there any good restaurants in the neighborhood
WITHOUT_SIM	√ do you have any good restaurants in the neighborhood
入力文	このカートは使えますか
PROPOSED	can I use this cart
WITHOUT_SIM	√ do you accept this cart
入力文	警察へ連絡してください
PROPOSED	please notify the police
WITHOUT_SIM	please contact the police
入力文	サイズがわかりません
PROPOSED	i 'm not sure of this size
WITHOUT_SIM	i do n't know this size
入力文	メニューをお願いします
PROPOSED	give me the menu please
WITHOUT_SIM	please the menu
入力文	アクセサリーはどこで買えますか
PROPOSED	√where can i buy clothes for accessory
WITHOUT_SIM	where can i get accessory
入力文	旅行の目的は何ですか。
PROPOSED	√what 's the purpose of the travel agency
WITHOUT_SIM	what 's the purpose of the trip

* 太字は PROPOSED と WITHOUT_SIM の翻訳が異なっている箇所を示す。√は誤りと評価された文を示す。下線はアライメント誤りを示す。

表 3 自動評価手法

BLEU	正解との n-gram の適合率の相乗 (幾何) 平均 (Papineni, Roukos, Ward, and Zhu 2002).
NIST	正解との n-gram の適合率の相加 (算術) 平均 (Doddington 2002).
WER	Word Error Rate. 正解との編集距離 (Niessen, F. J. Och, and Ney 2000).
PER	Position Independent Word Error Rate. 語順を用いない正解との編集距離 (Och, Uereng, and Ney 2001).
GTM	general text matcher. 正解と一致した最長語列の適合率, 再現率の調和平均 (Turian, Shen, and Melamed 2003).

* BLEU, NIST, GTM については大きな値ほど精度がよい。WER, PER については小さな値ほど精度がよい。

表 4 実験結果

	bleu	nist	wer	per	gtm
PROPOSED	0.41	8.04	0.52	0.44	0.67
WITHOUT_SIM	0.42	7.67	0.49	0.42	0.68
BASIC	0.39	7.92	0.52	0.44	0.67
BASELINE	0.31	6.65	0.62	0.54	0.59
C1	0.13	5.47	0.75	0.60	0.56
C2	0.27	7.31	0.54	0.47	0.65

表 5 誤り分析

数	分類	説明
21	DATA-SPARSENESS	用例の数が足りないことが原因である翻訳誤り. この場合, システムは翻訳辞書 (アライメントの際に用いた辞書) を用いて訳語を得るが, しばしば不適切な訳語が得られる)
6	ZERO-PRONOUN	ゼロ代名詞による翻訳誤り. ゼロ代名詞が入力文にだけ含まれており, 用例には含まれていない場合. または, 逆に, ゼロ代名詞が用例にだけ含まれて, 入力文には含まれていない場合に, 翻訳文から代名詞が抜け落ちてしまう.
4	ALIGNMENT-ERR	アライメントが不適切な用例を用いたことによる翻訳誤り.
3	WORD-ORDER	語順がおかしい場合.
3	SELECTION-ERR	適した用例が存在するが, それが選択されず, 不適切な用例を用いた場合.
12	OTHERS	上記のさまざまな誤りが複数存在し, 特定の原因に分類できない場合.

表 6 翻訳例

判定 誤り原因	入力文 システム出力
	市役所行きのバス停はどこですか。 where is the bus stop for the city hall だいぶよくなったみたい。 i feel much better
	インストラクターを紹介していただけますか。 would you please suggest a instructor 歩いてどのくらいですか。 how long does it take to walk please
	チェックインは何時からですか。 what is the check-in time
✓ ZERO-PRONOUN	ひどい頭痛がしているんです。 has a bad headache
	バスでそこへ行けますか。 can i get there by bus
✓ WORD-ORDER	あの棚の本を見たいのですが。 i 'd like to see the rack that are a books
✓ ZERO-PRONOUN	なにかメッセージが届いていませんか。 have received any message for something
✓ DATA-SPARSENESS	この綿のセーターを試着してもよろしいですか。 i would like to try this sweater for an cotton バス付きの二人部屋にします。 a two people room with a bath 日本円をドルに替えてください。 change our japanese yen please into dollars 切符売り場はどこでしょう。 where can i find the ticket office

* ✓は誤りと評価された文を示す.

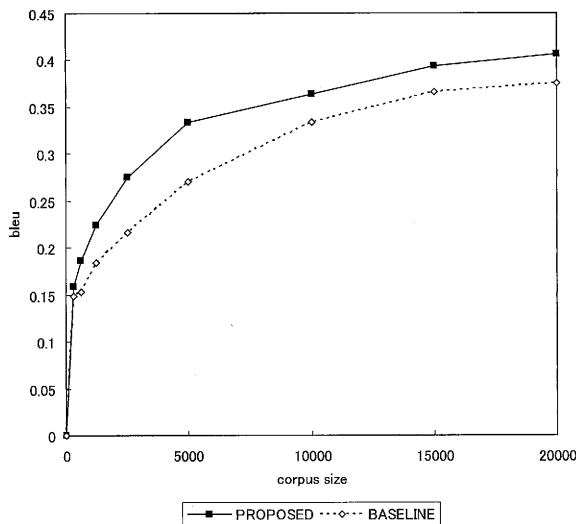


図 5 コーパスサイズと精度 (BLEU)

ら、もし、より多くのコーパスが利用可能であれば、精度はさらに向上すると考えられる。

また、次に ZERO-PRONOUN (ゼロ代名詞の問題) が多い。現在、提案手法はゼロ代名詞に関して、特別な処理を行っていないが、今後、省略解析の技術を用いて、より注意深くゼロ代名詞を扱うことが必要であろう。

参考までに、表 6 に翻訳例と分類結果の一部を示す。

4.5 コーパスサイズと精度

最後に、コーパスサイズ (トレーニングセットの対訳文数) と翻訳精度 (BLEU) の関係について調査した。これは、PROPOSED と BASELINE の 2 つのシステムを用いて行った (図 5)。

図 5 に示されるように、PROPOSED と BASELINE の差はコーパスサイズが小さい場合 ($x \approx 5,000$) に大きいことが分かる。このことから、PROPOSED の方が用例の不足に対してより頑健であることが分かる。

また、スコアは今回の実験の最大の用例数 ($x = 20,000$) で飽和していない。このことから、もし、より多くの用例を得ることができれば、より高い精度を得ることが期待される。

5 関連研究

これまで様々な用例ベース翻訳システムが提案されてきたが、それらは経験則に基づいて用

例を選択しており, 提案手法のような確率的な尺度に注意を払っていない。

例えば, (Richardson, Dolan, Menezes, and Corston-Oliver 2001) はマニュアルドメインの用例ベース翻訳システムを提案した。彼らのシステムは, 用例と入力文の間で一致する部分のサイズのみを用いて用例を選択している。

(Furuse and Iida 1994; Imamura 2002) は, 一致サイズとコンテキストの類似度の両方を用いて用例を選択している。(Aramaki, Kurohashi, Kashioka, and Tanaka 2003) は, それらに加え, さらにアライメントのもっともらしさを用いて用例を選択している。これらの手法では, 複数の尺度をどのような重みで考慮するか, という重み付けの問題が存在する。

6 おわりに

本稿では, 大きな用例ほど翻訳確率が高くなるという考えに基づき, 翻訳確率だけを用いて用例を選択する用例ベース翻訳手法を提案した。実験の結果は, 従来の経験則による用例選択を行うシステムよりも僅かに高い精度を得ることができ, 提案手法の妥当性を示している。

本研究により, これまで, 統計ベース翻訳と比べて不透明であった用例ベース翻訳のアルゴリズムを定式化することで, 今後より一層緻密な用例ベース翻訳の議論が可能になると考えている。

参考文献

- Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M., and Tsujii, J. (2004). “Overview of the IWSLT04 Evaluation Campaign.” In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT2004)*, pp. 1–12.
- Aramaki, E. and Kurohashi, S. (2004). “Example-based Machine Translation using Structural Translation Examples.” In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT2004)*, pp. 91–94.
- Aramaki, E., Kurohashi, S., Kashioka, H., and Tanaka, H. (2003). “Word Selection for EBMT based on Monolingual Similarity and Translation Confidence.” In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003) Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 57–64.
- Aramaki, E., Kurohashi, S., Sato, S., and Watanabe, H. (2001). “Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation.” In *Proceedings of MT Summit VIII*, pp. 27–32.
- Brown, P. F., Pietra, S. A. D., cent J. Della Pietra, V., and Mercer, R. L. (1993). “The

- Mathematics of Statistical Machine Translation: Parameter Estimation.” *Computational Linguistics*, **19** (2).
- Charniak, E. (2000). “A maximum-entropy-inspired parser.” In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL 2000)*, pp. 132–139.
- Doddington, G. (2002). “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.” In *Proceedings of the Human Language Technology conference (HLT2002)*, pp. 257–258.
- Furuse, O. and Iida, H. (1994). “Constituent Boundary Parsing for Example-Based Machine Translation.” In *Proceedings of the International Conference on Computational Linguistics (COLING1994)*, pp. 105–111.
- Imamura, K. (2002). “Application of Translation Knowledgeacquired by Hierarchical Phrase Alignment for Pattern-based MT.” In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation (TMI2002)*, pp. 74–84.
- Koehn, P., Och, F. J., and Marcu, D. (2003). “Statistical phrase-based translation.” In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003)*, pp. 48–54.
- Kurohashi, S. and Nagao, M. (1994). “A Syntactic Analysis Method of Long Japanese Sentences based on the Detection of Conjunctive Structures.” *Computational Linguistics*, **20** (4).
- Nagao, M. (1984). “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.” In *Elithorn, A. and Banerji, R. (eds.): Artificial and Human Intelligence*, pp. 173–180.
- Niessen, S., F. J. Och, G. L., and Ney, H. (2000). “An evaluation tool for machine translation: Fast evaluation for machine translation research.” In *Proceedings of the international conference on Language Resources and Evaluation (LREC2000)*, pp. 39–45.
- Och, F. J., Tillmann, C., and Ney, H. (1999). “Improved alignment models for statistical machine translation.” In *Proceedings of the Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP1999)*, pp. 20–28.
- Och, F. J., Uerang, N., and Ney, H. (2001). “An efficient a* search algorithm for statistical machine translation.” In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2001) Workshop on Data-Driven Machine Translation*, pp. 55–62.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). “BLEU: a Method for Automatic

- Evaluation of Machine Translation.” In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2002)*, pp. 311–318.
- Richardson, S. D., Dolan, W. B., Menezes, A., and Corston-Oliver, M. (2001). “Overcoming the customization bottleneck using example-based MT.” In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2001) Workshop on Data-Driven Methods in Machine Translation*, pp. 9–16.
- Turian, J. P., Shen, L., and Melamed, I. D. (2003). “Evaluation of machine translation and its evaluation.” In *Proceedings of MT Summit IX*, pp. 386–393.
- Watanabe, T., Sumita, E., and Okuno, H. G. (2003). “Chunk-based Statistical Translation.” In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2003)*, pp. 303–310.
- 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (1997). 日本語語彙大系. 岩波書店.

略歴

- 荒牧 英治: 1998年京都大学総合人間学部基礎科学科卒業. 2005年東京大学大学院情報理工学系研究科博士課程修了. 博士(情報理工学). 現在, 東京大学附属病院企画情報運営部特任助手. 機械翻訳, 医療情報の研究に従事.
- 黒橋 禎夫: 1989年 京都大学工学部電気工学第二学科卒業. 博士(工学). 1994年 同大学院博士課程修了. 京都大学工学部助手, 京都大学情報学研究科講師, 東京大学大学院情報理工学系研究科助教授を経て, 2006年京都大学情報学研究科教授, 現在に至る. 自然言語処理, 知識情報処理の研究に従事.
- 柏岡 秀紀: 1993年 大阪大学大学院基礎工学研究科博士後期課程修了. 博士(工学). 同年 ATR 音声翻訳通信研究所入社. 1998年 同研究所主任研究員(現 ATR 音声言語コミュニケーション研究所). 1999年 奈良先端科学技術大学院大学情報学研究科客員助教授. 主に自然言語処理, 機械翻訳の研究に従事.
- 加藤 直人: 1986年早稲田大学理工学部電気工学科卒業. 1988年同大学院修士課程修了. 同年日本放送協会(NHK)に入局. 現在, NHK放送技術研究所に勤務. この間, ATR 音声翻訳通信研究所, ATR 音声言語コミュニケーション研究所に出向. 博士(情報科学). 機械翻訳, 対話処理, 音声言語処理, 自動要約の研究に従事.

(2005年9月27日受付)

(2006年1月26日再受付)

(2006年1月27日採録)