確率的用例ベース翻訳の実現

荒牧英治

1　はじめに

近年，統計ベース翻訳 [4] や用例ベース翻訳 [5] 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よいと考えられてきた．なぜならば，用例のサイズ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．したがって，大きな用例が利用可能なドメイン，すなわち，特許翻訳のような類似文が多いドメインにおいて，用例ベース翻訳の可能性が注目されている．しかし，これまでの用例ベース翻訳システムは，用例のサイズ/類似度などを経験則による指標で計算してきたため，統計ベース翻訳システムに比べて，そのアルゴリムが不透明でアドホックであった．

提案手法は，翻訳確率という尺度のみを用いて用例選択を行う．提案する翻訳確率は，統計ベースのそれとはことなり，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て構築される．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．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，実験を行う．

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を生成することである．例えば，“彼は CD をかける” を翻訳する場合，“かける” 単独で用例を検索した場合，“bet” や “run” など不適切な訳語を選ぶ可能性がある．そこで，“CD をかける” など大きなサイズで用例が存在する場合は，可能な限り大きな用例を用い，正確な訳語 (“play”) を用いたい．提案手法は，この用例の大きさに対する選好を確率的に定式化して実現する．

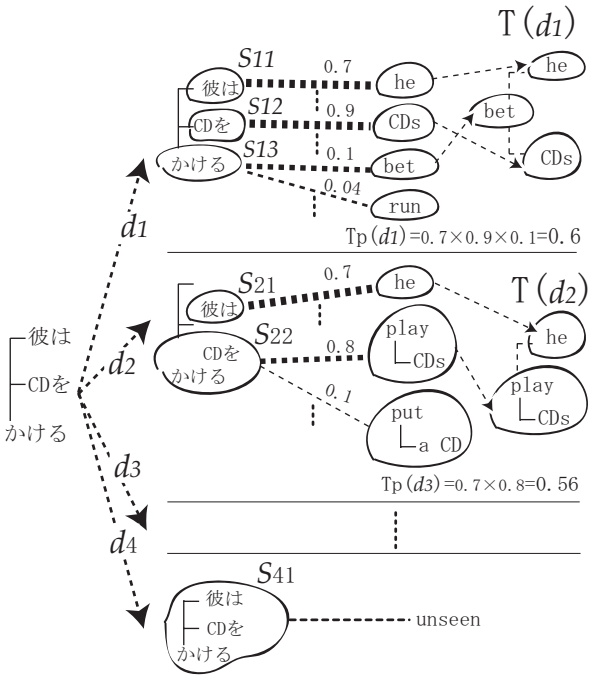

図 1: 翻訳のながれ．

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：

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

ここで，d_i は入力文の分解のパターン，D は d_i の集合とする．例えば，図 1 左の入力文の場合，$d_1, ..., d_4$ の 4 通りの部分木の組合せで表現できる．

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．

$$d_i = \{s_{i1}, s_{i2}, ..., s_{iM_i}\}, \tag{2}$$

ここで，s_{ij} は入力文の部分木である．例えば，図 1 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．

次に，各部分木 s_{ij} について, もっとも翻訳確率 $P(t_{ij} \mid s_{ij})$(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 $P(d_i)$ とする:

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

ここで，$t_{i1}, ..., t_{iMi}$ を d_i の翻訳と考え，$T(d_i)$ と表記する．

最後に，もっとも高い翻訳確率を持つ 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：

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

例えば，図 1 の $T(d_1)$ のように，入力文を小さな部分木に分解した場合は，曖昧性のある日本語 “かける” に対して，“bet”,“run” や “play” など様々な英語表現が考えられる．この場合，適切な訳である $P(play \mid \text{かける})$ 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，より大きな用例 “CD をかける” を用いている．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，大きな用例を検索した場合は，コーパス中に存在せず，確率が定義されない場合がある．

最後に，用例に翻訳確率について述べる．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る，この翻訳確率 $P(t \mid s)$ を次のように定義する:

$$P(t \mid s) = \frac{count(t, s)}{count(*, s)}, \quad (5)$$

ここで，$count(t, s)$ 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，$count(*, s)$ は日本語部分木 (s) の出現頻度である．

3 実験

実験は，(1) 提案システム (PROPOSED)，および，(2) 経験則によるメジャーにより用例を選択するシステム [2](BASIC) の 2 つの翻訳システムを自動評価法 (BLEU, NIST など) を用いて比較することで行った．

表 1: 実験結果.

	bleu	nist	wer	per	gtm
PROPOSED	0.41	8.04	0.52	0.44	0.67
BASIC	0.39	7.92	0.52	0.44	0.67

コーパスは IWSLT04[1] にて配布されたコーパス (トレーニングとテストセット) を用いた．トレーニングセットは旅行対話ドメインの 20k の日英対訳文からなる．これらに対して，翻訳辞書を用いた手法 [3] でアライメントを行った．推定された対応関係から用例の翻訳確率は計算した．利用した辞書は英辞郎などいくつか辞書をマージしたもので，延べ百万語を含む．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．

各手法の精度を表 1 に示す．表 1 に示されるように，提案手法 PROPOSED は，経験則による RULE 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．

4 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システムを提案した．

参考文献

[1] Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. Overview of the IWSLT04 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 1–12, 2004.

[2] Eiji Aramaki and Sadao Kurohashi. Example-based machine translation using structual translation examples. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 91–94, 2004.

[3] Eiji Aramaki, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe. Finding translation correspondences from parallel parsed corpus for example-based translation. In *Proceedings of MT Summit VIII*, pp. 27–32, 2001.

[4] Peter F. Brown, Stephen A. Della Pietra, Vi cent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, 1993.

[5] Makoto Nagao. 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, 1984.