2005－NL－165 (14)
2005/1/12

固有表現抽出器を用いた、非直訳文書対からの固有表現翻訳対獲得

熊野 正[†]
kumano.t-eq@nhk.or.jp

柏岡 秀紀[‡]
hideki.kashioka@atr.jp

田中 英輝[†]
tanaka.h-ja@nhk.or.jp

福島 孝博[§]
fukusima@res.otemon.ac.jp

[†] NHK 放送技術研究所　〒157-8510 東京都世田谷区砧 1-10-11
[‡] ATR 音声言語コミュニケーション研究所　[§] 追手門学院大学

概要

対訳ニュース原稿のような content-aligned な対訳文書から固有表現等の対訳を網羅的に獲得するための手法として先に提案した手法の改良として、統計翻訳における翻訳モデルを参考に構築した新たな統計的固有表現対応づけモデルを提案する。本手法は、各言語の文書中の固有表現を単言語固有表現抽出器を用いて網羅的に発見した後、同一対象を指すものどうしを各言語内でグループ化し、最後に今回提案する手法でグループ間に対応づけを行う。学習コーパスから獲得した固有表現翻訳確率を組み込んだ対応づけモデルを用いることで、グループ対応性能 F = 0.82 を達成した。

キーワード:　対訳抽出、content-aligned な対訳文書、固有表現、アラインメント

Acquiring Named Entity Translation Pairs from Content-aligned Corpora by Using NE Extraction Technique

Tadashi Kumano[†]
kumano.t-eq@nhk.or.jp

Hideki Kashioka[‡]
hideki.kashioka@atr.jp

Hideki Tanaka[†]
tanaka.h-ja@nhk.or.jp

Takahiro Fukusima[§]
fukusima@res.otemon.ac.jp

[†] NHK Science and Technical Research Laboratories
1-10-11, Kinuta, Setagaya-ku, Tokyo, 157-8510 JAPAN
[‡] ATR Spoken Language Translation Research Laboratories
[§] Otemon Gakuin University

Abstract

We propose a statistical alignment model of bilingual named entities (NEs) as a improvement of the previous method that acquires NE translation pairs from content-aligned corpora such as bilingual news articles. The method firstly extracts NEs from each of the bilingual document exhaustively by using NE extraction technique, then composes NE groups within each documents that share the same referent. Finally it corresponds NE groups between documents to extract bilingual NE translation pairs by applying the proposed alignment model. It achieved F = 0.82 as the corresponding performance of NE groups, when the alignment model was enhanced to consider lexical translation probabilities derived from the training data.

Keywords:　translation pairs acquisition, content-aligned bilingual documents, named entities (NEs), alignment

1　はじめに

人名、地名、組織名といった、いわゆる固有表現（Named Entity; NE）は、文書の具体的内容を伝達する、重要な役割を担っている。例えば、ニュースのような具体的事実を伝達する意図で書かれた文書を翻訳するとき、固有表現が適切に理解されるよう翻訳しなければ、他の部分の翻訳がいかに適切であって

も、文書全体の役割が損なわれてしまう。固有表現の翻訳には原則として類推が働かないため、ある固有表現の翻訳するにはとにかくその対訳を知っていなければならない。ところが、固有表現は具象を指し示す名称であり、刻一刻と新出する性質を持っているため、どれだけの数の固有表現に対してその対訳を知っていてもすぐに陳腐化し、知識不足が顕著になってくる。それゆえ、機械翻訳システムを維持していくにあたって、固有表現の新出に追従して対訳知識を拡充していくことは、非常にコストがかかるが欠かすことのできない作業である。

このような固有表現対訳の拡充を自動的に行う手法として、先に我々は、対訳ニュース記事のような固有表現を多く含む文書対から網羅的に固有表現対訳対を獲得する手法を提案した [3]。我々が対象とするこのような対訳文書対は、文書全体としては同一の話題を伝達する意図で書かれたものだが、一般に逐語訳からは離れており、文などの単位での対応が見いだせるとは限らない。また、固有表現の性質上、個々の表現の文書中での出現頻度は一般に小さく、大規模コーパス中にただ 1 度しか現われないものも多い。そのため、従来対訳表現対を獲得するのによく用いられてきた、共起頻度に基づく手法や構造的な対応を手掛かりとする手法はうまく機能しないことが予想される。これを受けて我々は、固有表現抽出技術を用いて各言語文書内にある抽出対象表現を網羅的に列挙したうえ、文書中に出現する固有表現の分布（何について、どのような順序で提示されているか）が対訳言語間で類似している性質を利用して両者を結びつけることで、ある一定の性能を達成できることを示した。この性質は、放送ニュース原稿からなる NHK 日英ニュースコーパスに人手で対訳固有表現タグを付与したものを分析した結果 [2] によるものだが、これは、我々が分析した文書に特有の性質ではなく、同一内容を伝達しようとする（”content-aligned”）対訳文書に共通の性質であると考える。

先に提案した手法では、最良の対応づけを選択するのに必要な、対応づけのよさを測る尺度として、ad hoc な指標を用いていた。しかし、この対応づけ指標が対象文書対における対訳固有表現の振舞いに適合したものであるかどうかには疑問の余地があった。本稿では、この問題点の解決策として、統計翻訳 [1] における翻訳モデルに着想を得た統計な日英固有表現間の対応づけモデルを導入した結果について報告する。2 節では、本手法が解決する問題の定義を説明する。3 節では今回提案する手法の説明として、まず先の手法と共有する部分について簡単に紹介した後、今回改良を行った対応づけモデルについて、その詳細を説明する。そして 4 節で提案手法の性能評価を行い、最後に 5 節でまとめと今後の展望について述べる。

2 問題設定

content-aligned な対訳文書においては、ある共通の具体的事象を指し示す固有表現が各々に複数存在する場合、それらのどれとどれが実際に対訳関係にあるのかを決定することが可能とは限らない。そこで、このような対訳文書から固有表現の対訳対を抽出する問題を、対訳文書に以下の 3 種類の情報を付与するものと定義する（図 1）[3]。

1. 各言語文書中の固有表現の出現と、その各々の固有表現種別（人名、地名など）
2. 各言語文書内で、同一の具体的対象を指し示す、同一固有表現種別の固有表現グループ
3. 同一固有表現種別に属する固有表現グループ間の 1 対 1 対応

この結果得られる固有表現グループ対の要素である各固有表現の全ての可能な組み合わせを取り出すことで、対訳辞書を獲得することができる。

我々は、この設定に基づいて、NHK 日英ニュース原稿 2,000 記事対に人手でこれら 3 種類の情報を付与したコーパス（以後「タグつき日英ニュースコーパス」と呼ぶ）を構築し、分析を行っている [2]。

3 提案手法

先に提案した手法 [3] と同様に、2 節で定義した 3 種類の情報を順番に、3 つの別個の処理: 1. **固有表現抽出**、2. **言語内グループ化**、3. **言語間対応づけ** にて付与する。今回提案する手法においては、上記 1. と 2. は先の提案のものをそのまま利用する。以下では、1., 2. についてごく簡単に紹介した後、3. の統計的対応づけ手法について説明する。

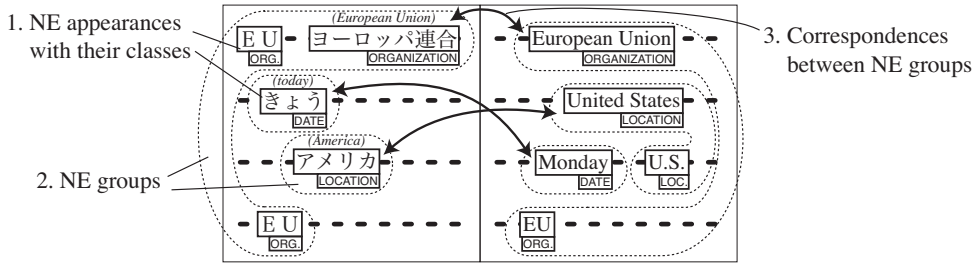

図 1　対訳文書間における固有表現の対訳関係の定義

3.1　固有表現抽出

山田ら [5] が提案した、Support Vector Machine（SVM）を用いた日英の固有表現抽出器を、汎用チャンカ YamCha[1] を利用して実装した。「タグつき日英ニュースコーパス」は、日本語、英語とも IREX 固有表現抽出タスク [4] 仕様に準拠したタグを付与してあり、これを用いて学習した日英の固有表現抽出器は、どちらも IREX 仕様の 8 種の固有表現種別（組織名、人名、地名、固有物名、日付、時間、金額、割合）の固有表現や数値表現（以後「固有表現等」と総称する）を抽出する。

抽出精度は、日本語で F = 約 0.94、英語で F = 約 0.92 程度である。

3.2　言語内グループ化

以下の基準に合致する、同一固有表現種別に属する 2 つの固有表現を同一グループと見なす処理を繰り返すことで、グループ化を行う。

日本語:

- 一方を構成する全ての文字が他方にその順序で含まれている
- 数値の表記を正規化語同一となる

英語:

- 一方を構成する全ての単語が他方にその順序で含まれている
- 一方が他方の頭字語となっている

グループ化性能は、日本語で F = 約.66、英語で F = 約.82 程度である。

3.3　統計的手法を用いた言語間対応づけ

言語間対応づけ処理は、固有表現グループ情報が付与された対訳文書対に対し、可能なグループ間 1 対 1 対応づけ集合の中から最良のものを選択する課題である。

タグつき日英ニュースコーパスの分析結果より、固有表現グループが相手言語文書内に対応する（同一対象を指し示す）固有表現グループを持つ割合（訳出率）や、文書内の 2 つのグループ間の出現順序を各々の要素のうち最も文書の先頭に近いもの同士の順序と定義したときに対応先を持つ 2 つのグループの文書中での出現順序と相手言語文書中での各々の対応先グループの出現順序とが一致する割合（順序保存率）が高い対応づけ集合の方が、よりもっともらしい対応づけである傾向がある。先の提案手法では、上記 2 つの指標の線形結合にて対応づけ集合の評価を行った。

しかし、我々が扱っている文書対は逐語訳ではないため、一方言語側に現われた固有表現グループが相手言語側に完全に訳出され、かつ順序も完全に保存されている、ということはまれである。従って、いくつかの非訳出や順序非保存の変則を含む対応づけ集合どうしを比較する際には、個々の変則が「この種の対訳コーパス」の性質から見てどれだけ「あり得る」かを細かく反映できるような尺度を導入することが望ましい。

本節では、統計翻訳の翻訳モデルの 1 つである IBM model 3 [1] を参考に、尺度の統計モデル化を行う。具体的には、原言語文書中の固有表現グループの出現系列から相手言語文書中の固有表現グループ出現系列を生成する確率をモデル化し、それによって個々の対応づけ集合の尤度を求める。以下にまず基本となるモデルを提案し、その後に、対訳コーパスの性質をより細かく反映するためのモデルの精緻化をいくつか示す。

[1] http://chasen.org/~taku/software/yamcha/

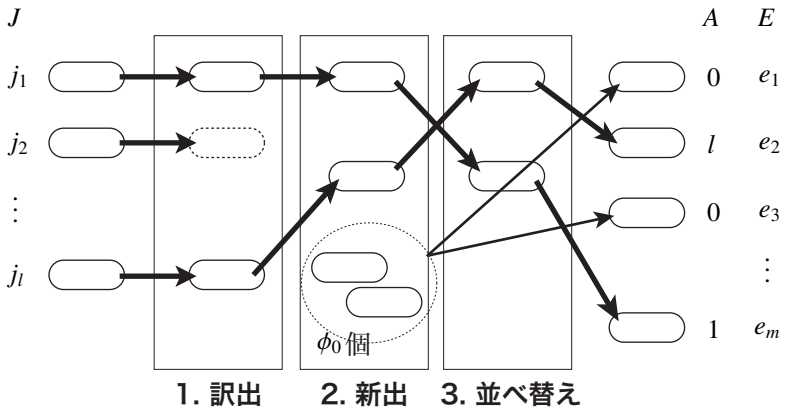

図 2　原言語固有表現系列から相手言語系列を生成するモデル

3.4　基本モデル

原言語文書中の l 個の固有表現グループ $\{j_1, j_2, \ldots, j_l\}$、および相手言語文書中の m 個の固有表現グループ $\{e_1, e_2, \ldots, e_m\}$ の各々は、固有表現種別によってクラス化されている。すなわち、固有表現種別集合を C (∋ ex. 人名, 地名, ...) として、

$$j_i \in C, e_k \in C \quad (1 \leq i \leq l, 1 \leq k \leq m) \tag{1}$$

原言語文書中の固有表現グループを出現順序（各グループの先頭要素の順）に沿って並べた系列 $J = j_1 j_2 \ldots j_l$ から、相手言語文書中の固有表現グループを同様に並べた系列 $E = e_1 e_2 \ldots e_m$ を生成するモデルを考える。このモデルは、J から E への変換の過程に、図 2 のように、原言語固有表現グループの 1. **訳出**、相手言語固有表現グループの 2. **新出**、そして固有表現グループの 3. **並べ替え**の 3 つの段階を想定する。

3.4.1　訳出モデル

原言語固有表現グループの各々について、相手言語側固有表現グループに訳出するかしないか（言い換えると、原言語固有表現グループの各々が、相手言語側に対応する固有表現グループを持つかどうか）を決定する。

原言語側の系列 J 中の固有表現グループ j_i $(1 \leq i \leq l)$ を相手言語側に訳出するかどうかを、ϕ_i $(\in \Phi)$ で表すことにする。

$$\phi_i = \begin{cases} 0 \cdots \text{訳出しない} \\ 1 \cdots \text{訳出する} \end{cases} \tag{2}$$

このとき、系列 J 下で訳出系列 Φ が生起する確率（訳出尤度）を以下のように求める。

$$P(\Phi|J) = \prod_{i=1}^{l} r(\phi_i|j_i) \tag{3}$$

ここで r を「訳出率」と呼ぶ。基本モデルでは訳出率 r を j によらず一様として扱う。

IBM Model 3 における fertility model に相当するが、本モデルにおいてはグループ間の対応は 1 対 1 に限られるので、ϕ_i は 2 以上にならない。

3.4.2　新出モデル

原言語文書中の固有表現グループから訳出されたものでない、新出の固有表現グループを生成する。

新出の固有表現グループの個数を ϕ_0 個とする。ϕ_0 個の新出グループの集合 $N = \{n_1, n_2, \ldots, n_{\phi_0}\}$ を生成する確率（新出尤度）を以下のように求める。

$$P(N) = \prod_{i=1}^{l} g(n_i) \tag{4}$$

ここで g を「新出率」と呼ぶ。基本モデルでは新出率 g を n によらず一様として扱う。

IBM Model 3 における NULL generation model に相当する。しかし、NULL generation model は原言語単語列の単語間に NULL を挿入するモデルであるが、本モデルでは新出グループは原言語のグループ系列とは別個に生成し、また新出グループどうしの間に順序を仮定しない。これは、次の段階の並べ替えモデルにおいては原言語から訳出されたグループどうしの間の順序のみを考慮の対象とし、原言語から訳出されたグループと新出グループとの間に制約をおかないため、本モデル中で新出モデルの順序を

決定する必要がないためである。

3.4.3 並べ替えモデル

並べ替えに先立ち、原言語から訳出された固有表現グループと新出グループの各々を、相当する相手言語の固有表現グループと読み替える（後述の記法を用いるならば、$P(E|A) \equiv 1$）。IBM Model 3 における lexicon model に相当するが、本基本モデルにおいては訳出関係にある原言語と相手言語の固有表現グループは固有表現種別集合 C の同一要素でクラス化されたものであるため、実際の操作は発生しない。

並べ替えモデルは、原言語固有表現グループに由来する固有表現グループの順序を並べ替え、その間に適宜新出固有表現グループを配置して最終的な系列 E を生成する。

系列 E の各要素 e_k $(1 \leq k \leq m)$ が由来する原言語固有表現グループの、系列 J 中での位置を a_k $(\in A)$ で表すことにする。

$$a_k = \begin{cases} 0 \cdots \text{新出グループ} \\ i \cdots j_i \ (1 \leq i \leq l) \text{ に由来} \end{cases} \tag{5}$$

このとき、系列 J、訳出 Φ、新出グループ N 下で対応 A が生起する確率（並べ替え尤度）を以下のように求める。

$$P(A|J,\Phi,N) = \prod_{p=1:a_p\neq 0}^{m-1} \prod_{q=p+1:a_q\neq 0}^{m} d(sgn(a_q - a_p)|j_{a_p}, j_{a_q}, e_p, e_q) \times \frac{1}{{}_mC_{\phi_0}} \tag{6}$$

ここで d を「順序保存率」と呼ぶ。順序保存率は、原言語由来のある 2 つの相手言語固有表現グループ e_p, e_q と各々の由来である原言語固有表現グループ j_{a_p}, j_{a_q} が、各々の属する系列内での出現位置の前後関係が一致する（$a_q - a_p > 0$ となる）割合を表す。基本モデルでは順序保存率 d を $j_{a_p}, j_{a_q}, e_p, e_q$ によらず一様として扱う。また、最後の ${}_mC_{\phi_0}$ による補正は、原言語に由来する $m - \phi_0$ 個のグループの並べ替え結果をその順序関係を壊さないように m 個の系列中に展開するやり方が ${}_mC_{m-\phi_0} = {}_mC_{\phi_0}$ 通りあり、A はその中の 1 通りを選択していることによる。

並べ替えモデルは IBM Model 3 における distortion model に相当するが、先に述べたように、並べ替えは原言語に由来する固有表現グループどうしの間にのみ定義する。また、並べ替え尤度の定式化に、先に提案した手法でも用いた「順序保存率」を用いた点も異なっている。

最後に、新出グループを $a_k = 0$ である各 e_k に割り当て、系列 E を完成する。割り当て方は $\phi_0!$ 通りあるから、最終的な A, E の生成確率（すなわち対応尤度）は次の式となる。

$$\begin{aligned} &P(A,E|J) \\ &\quad = P(\Phi|J) \cdot P(N) \cdot P(A|J,\Phi,N) \cdot P(E|A) \cdot \frac{1}{\phi_0!} \\ &\quad = P(\Phi|J) \cdot P(N) \cdot P(A|J,\Phi,N) \cdot \frac{1}{\phi_0!} \end{aligned} \tag{7}$$

3.5 訳出・新出モデルの精緻化

基本モデルにおいて、訳出モデルで用いる訳出率 r と新出モデルで用いる新出率 g は、それぞれ j や n の種類によらず一様として扱った。これらをより細かく分類して異なる確率を与えることで、訳出モデルや新出モデルの精緻化を図ることができる。

具体的には、次の拡張を検討した。

a-1. 固有表現種別による分類
　我々の分析した日英コーパスでは、例えば日本語文書中の時間表現は他に比べて顕著に訳出率が低い、といった、固有表現種別によって異なる傾向が見られた。

a-2. グループ要素の個数による分類
　固有表現グループの要素の多さは、文書中で繰り返し言及されていることを意味している。このような表現は重要な対象を指しており、訳出される可能性が高いことが推測できる。

3.6 並べ替えモデルの精緻化

同様に、基本モデルにおいて、並べ替えモデルで用いる順序保存率 d は、j や e の種類によらず一様として扱った。これらをより細かく分類して異なる確率を与えることで、並べ替えモデルの精緻化を図ることができる。

具体的には、次の拡張を検討した。

b-1. 原言語でのグループ間距離による分類
　原言語側の 2 つのグループ j_{a_p}, j_{a_q} が離れていればいるほど、その順序は各々の相手言語側での訳出先 e_p, e_q でも保存されやすい。

b-2. 訳出先が同一文内である場合の特例
原言語側の 2 つのグループの訳出先が相手言語側で同一文内となる場合、とりわけ日→英など構文構造に大きな違いのある言語対においては、構文上の理由で順序の保存性が著しく損われる場合がある。

b-3. 固有表現分類種別が同一かどうかによる分類
我々の分析した日英コーパスでは、固有表現分類種別が同一であるグループどうしの順序保存率は、そうでないものに比べて顕著に高かった。

上記 b-2. のような、訳出先がそれぞれ同一文内かどうかを考慮するモデルを採用すると、(6) 式の最後の補正項を変更する必要が生じる。相手言語側文書が s 文からなっており、並べ替えの結果、原言語に由来する固有表現グループが t 文に分散するとすると、補正項の分母は次式のようになる。

$$ {}_sC_t \cdot {}_{m+t-1}C_{\phi_0} \tag{8} $$

3.7 語彙モデルの導入

これまでのモデルにおいては、原言語固有表現グループ j や相手言語固有表現グループ e がどのような具体的な固有表現等を含んでいるかは考慮せずに固有表現種別のみを参照して対応づけを行っており、IBM Model 3 における lexicon model に相当するものを導入しなかった（$P(E|A) \equiv 1$）。これは、固有表現はその性質上、大規模コーパスにおいても出現頻度が小さいためであり、手法の頑強性に寄与している。

しかし、やはりこれまでのモデルで用いた情報のみで完全な対応づけを実現するのは難しい。そこで、ある程度の規模の学習コーパスから固有表現等（グループではない）の翻訳確率を獲得できていることを前提に、語彙モデルを導入することを検討する。

前提として、あらかじめ用意した固有表現等翻訳確率は、学習コーパスに現われなかった未出現表現に関わる確率として妥当な値を与えられるよう、適宜スムージングを行っておく必要がある。

語彙モデルは、原言語から訳出された固有表現グループの各々について、その要素である原言語固有表現等の集合から、対訳として相手言語固有表現等の集合を生成する。

まず、原言語固有表現等の集合 $V = \{v_1, v_2, \ldots, v_u\}$ からなる原言語固有表現グループから相手言語固有表現等 の集合 $W = \{w_1, w_2, \ldots, w_x\}$ からなる相手言語固有表現グループを生成する確率 $P(W|V)$ は、以下のように求めることができる。

$$ \begin{aligned} &P(W|V) = \\ &P(w_1 \cup w_2 \cup \cdots \cup w_x \,|\, v_1 \cup v_2 \cup \cdots \cup v_u) = \\ &\begin{cases} \dfrac{\displaystyle\sum_{i=1}^{u}\sum_{k=1}^{x} P(w_k|v_i)P_{ML}(v_i)}{\displaystyle\sum_{i=1}^{u} P_{ML}(v_i)} & \left(\displaystyle\sum_{i=1}^{u} P_{ML}(v_i) > 0\right) \\ \displaystyle\sum_{i=1}^{u}\sum_{k=1}^{x} P(w_k|v_i)P(w_k|v_i) & \left(\displaystyle\sum_{i=1}^{u} P_{ML}(v_i) = 0\right) \end{cases} \end{aligned} \tag{9} $$

ここで、$P_{ML}(v)$ は学習コーパスから最尤推定した v の出現確率、$P(w|v)$ は同じコーパスから学習し適切にスムージングした $w \to v$ の翻訳確率を表す。また、新出グループ n から W を生成する確率 $P(W|\emptyset)$ には、n から生成可能な全ての固有表現等の異なり数を $N(n)$ として、一律に 0 グラム確率 $1/N(n)$ を与えておく。

これを用い、各 e_k $(1 \le k \le m)$ に対して由来する原言語固有表現グループの要素集合 $V(a_k)$ $(\in \boldsymbol{V})$ から生成した相手言語側要素集合 $W(k)$ からなる系列 $\boldsymbol{W}$ を生成する確率（語彙尤度）は次のようになる。

$$ P(\boldsymbol{W}|\boldsymbol{V}, A) = \prod_{k=1}^{m} P(W(k)|V(a_k)) \tag{10} $$

最終的に、語彙モデルを導入した対応尤度は次の式となる。

$$ P(A, E, \boldsymbol{W}|J, \boldsymbol{V}) = P(\boldsymbol{W}|\boldsymbol{V}, A) \cdot P(A, E|J) \tag{11} $$

4 実験

本節では、「タグつき日英ニュースコーパス」に対し、固有表現の出現とグループ情報はタグの情報をそのまま用いてグループ間対応づけ情報を提案手法で付与する実験を行った結果を報告する。

4.1 実験条件

実験は、「タグつき日英ニュースコーパス」2,000 記事対を学習データと評価データとに分割（10 分割交差検定、学習データ:評価データ = 9 : 1）して、学習データから対応づけに必要な統計量を学習し、評価データに対して適応する形で行った。

対応づけのための対応尤度モデルとしては、前節で提案した、基本モデル、およびこれに対して 3.5〜3.6 節で検討した拡張のいくつかを加えたものを数種類用意し、性能を比較する。具体的には、用いたモデルは以下の通りである。

1. 基本モデル
2. 1. + 拡張 a-1
   （訳出・新出率を固有表現種別により細分化）
3. 2. + 拡張 a-2
   （訳出・新出率をグループ要素数により細分化）
4. 1. + 拡張 b-1
   （順序保存率を原言語でのグループ間距離（文数）により細分化）
5. 4. + 拡張 b-2
   （順序保存率を訳出先が同一文内であるかにより細分化）
6. 5. + 拡張 b-3
   （順序保存率を固有表現種別が同一であるかにより細分化）
7. 3. + 6.
8. 7. + 語彙モデルの導入
9. 語彙モデルのみ（比較のため）

語彙モデルの導入にあたって学習コーパスから獲得する必要がある、固有表現等の翻訳確率 $P(v_E|v_J)$ の計算時には、Katz のスムージング法を用いて頻度 5 以下の共起頻度のディスカウントを行い、かつディスカウントされた確率を未出現共起へ配分する時にはその 1/2 を未出現語に関わる確率値に割り当てた。

4.2 結果

各モデルを用いて最尤であると判断した対応づけ集合が、正解データの対応づけ集合に含まれる個々の対応関係をどれだけ復元したかを評価した結果を、表 1 に示す。表中、モデル E_{CO}, $E_{CO'}$ はどちらも先の提案手法 [3] での対応づけ性能である。手法の詳細については割愛するが、用いている情報はおおよそ、E_{CO} が「基本モデル + 拡張 a-2」、$E_{CO'}$ が「E_{CO} + 拡張 b-1′（同一文かどうかのみ）+ 拡張 b-2」、に相当する。

実験結果から、本稿で提案した統計に基づく対応づけのモデル化は一定の成果を挙げ、先に提案した手法と同程度の性能を達成している。また、モデル 8. の性能評価から、語彙モデルの導入は本タスクの性能向上に非常に有効であったと言える。モデル 8. とモデル 9. の性能比較から、提案手法の基本的な枠組みと語彙モデルは相補的に働くと考えられる。

表 1　実験結果

モデル番号	精度	再現率	F 値
1.	.553	.563	.554
2.	.560	.589	.571
3.	.582	.611	.592
4.	.557	.575	.562
5.	.569	.537	.547
6.	.569	.539	.548
7.	.609	.607	.603
8.	.817	.835	.821
9.	.480	.492	.487
E_{CO}	.548	.600	.569
$E_{CO'}$	.582	.637	.605

現時点では、今回提案した統計的モデル化の先の提案手法に対する性能面での優位性を検証することができなかった。本稿で提案したモデルのうち、並べ替えモデルにおける並べ替え尤度の計算を訳出される固有表現グループの全ての組み合わせについて順序保存率を積算することで行っているが、これがあまり確率のよい近似になっていないのではないかという感触を持っている。引き続き検討を行いたい。

5 まとめ

対訳ニュース原稿のような content-aligned な対訳文書から固有表現等の対訳を網羅的に獲得するための手法として先に提案した手法を改良し、統計翻訳における翻訳モデルを参考に新たな統計的固有表現対応づけモデルを導入した結果を報告した。また、学習コーパスから獲得した固有表現翻訳確率を対応づけモデルに組み込むことで、顕著な性能向上が得られた。現在のところ、今回の提案手法は先の提案に対して対応づけ性能の優位性を検証することができなかったが、今後の検討によってさらなる性能向上が期待できる。

今後は、提案手法による固有表現翻訳対獲得の成果を機械翻訳や翻訳者支援に適用していくつもりで

ある。また、従来の手法とは相補的な文書間対応づけの指標として、より広範な表現翻訳対の獲得の手がかりとならないか、あるいは統計翻訳モデルへの適用ができないかを検討したい。

謝辞

本研究は、その一部を、独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Compuatational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[2] Tadashi Kumano, Hideki Kashioka, Hideki Tanaka, and Takahiro Fukusima. Construction and analysis of Japanese-English broadcast news corpus with named entity tags. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition: Combining Statistical and Symbolic Models*, pp. 17–24, 2003.

[3] Tadashi Kumano, Hideki Kashioka, Hideki Tanaka, and Takahiro Fukusima. Acquiring bilingual named entity translations from content-aligned corpora. In *Proceedings of the 1st Internationa Joint Conference on Natural Language Processing (IJCNLP-04)*, pp. 270–277, 2004.

[4] 関根聡, 井佐原均. IREX プロジェクト概要. IREX ワークショップ予稿集, pp. 1–5, 1999.

[5] 山田寛康, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.