確率手法に基づくニュース文書からの領域依存情報抽出に関する研究

揚石, 亮平 / AGEISHI, Ryohei

(発行年 / Year)
2010-03-24
(学位授与年月日 / Date of Granted)
2010-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2009年度　修 士 論 文

確率手法に基づくニュース文書からの領域依存情報抽出に関する研究

STUDIES ON DOMAIN-DEPENDENT INFORMATION EXTRACTION FROM NEWS DOCUMENTS BASED ON PROBABILISTIC METHOD

指導教官　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

08R3101　揚石 亮平
Ryohei AGEISHI

目 次

第1章　序論

1.1　問題の背景

近年のインターネット，Web 技術の発達に伴ない，多種多様なデータを誰もが容易に入手し利用できるようになった．これらのデータは日本語だけに留まらず，多言語による情報も容易に入手することが可能になっている．しかし，これらのデータはタグやコマンドなどの構造的手がかりを有さないデータとして蓄積されている．膨大なデータから利用者が一つ一つのデータ内容を吟味し，重要な情報や知識を獲得するのは到底不可能であるため，計算機による支援が急務となっている．特に現在では，日本国内の情報だけでなく世界中の情報を得なければならない機会が多いことから機械翻訳に大きな注目が集まっている．

1.2　統計機械翻訳

機械翻訳には大きく分けて 2 つの手法が存在する．1 つは翻訳のための対訳辞書や文法規則を人手で記述していく，**文法ベース**による手法である．この手法の特徴は，人間が英日翻訳を考える際の，1) 英文の意味を考える，2) 同等な意味の日本語文を出力する，というステップを, 1) 英文の意味構造を解析し，2) 得られた意味構造から日本語文を出力する，として実現したことである．文法ベースの手法は，適切な文法を与えれば，文法的に正しい翻訳文を生成できる．しかし，すべての文法を記述することは不可能だという問題が存在する．もう 1 つは，対訳コーパスという人間による翻訳サンプルから確率的に翻訳を行う**統計機械翻訳** (Statistical Machine Translation) による手法である．統計機械翻訳手法の特徴は，英文 e が日本語文 j に翻訳される確率 $P(j|e)$ が最大になるような日本語文 j を出力することである．統計機械翻訳は，意味構造や文法規則を全く用いず，対訳コーパスから翻訳確率を学習しているため，日本語として自然な文を出力することが可能である．アルゴリズムが言語に依存しないため，学習データさえあれば容易に多言語化が行えるという利点を持つ．現在では，利用可能な対訳コーパスが大量にあることから統計機械翻訳による手法が機械翻訳分野の大半を占めている．

しかし，統計機械翻訳手法には，単語や言語構造が似ている欧米の言語では効果的に働くが，単語や言語構造が大きく異なる英語日本語間では翻訳が困難であるという問題がある．この問題は単語や言語構造が異なるため，$P(j|e)$ を効果的に学習できな

いことに起因している．したがって単語や言語構造の違いを扱うための意味・文法情報を用いる必要がある．また，統計翻訳では頻度を用いて学習を行っているため，中・低頻度の確率は無視されてしまいがちである．しかし，特定領域においては中・低頻度のものが重要となる場合がある．例えば，*book*は**本**または**予約する**に翻訳される可能性が高いが，スポーツ分野においては**警告を出す**という意味で用いられることが多い．このような意味・文法情報や領域に依存した情報を統計機械翻訳の手法に付加することで翻訳精度を上昇させることが可能である．そのため，学習データからこのような情報を抽出するための手法を確立することが急務となっている．

1.3 扱う問題

本研究では統計機械翻訳の手法に付加する意味や文法情報を抽出する手法のうち，

- 形態素解析
- アラインメント
- 類義語

の 3 つに着目して論じる．

形態素とは，これ以上に細かくすると意味を失う最小の文字列 (単語) 情報をいう．自然言語の文章に対して，各文の単語識別 (トークン化)，その語形変化 (語幹抽出や語尾変化)，品詞同定処理を総称して形態素解析という．形態素解析により単語の多義性を排除できる可能性がある．例えば，*book*という単語の品詞を特定できれば，*book*が**本**と**予約する**のどちらの意味を表現しているかを決定できる．特に固有名詞は文の中で重要な役割を果たしているため正確に推定する必要がある．

次に，アラインメント問題を考える．アラインメントは英日言語間の単語がどのように対応しているかを確率的に表現したものである．したがってアラインメントを正確に行うことで，文法構造が大きく異なる言語間においても単語の対応を取ることが可能となり，高精度の翻訳の実現が可能となると考えられる．

最後に，類義語問題を考える．人間が翻訳文を判断する場合，人間は単語の意味を考えることが可能なため，対訳文の意味が類似していれば翻訳文は正しいと判定することができる．しかし，計算機は意味を扱うことができないため，単語が異なっていれば意味が同じであっても誤りであると判定してしまう．ここに類義語を用いれば，計算機が意味を考慮して翻訳文の判断を行えるかもしれない．特に，**立つ**と**出馬する**がニュースという領域で依存するように，特定の領域で類似するという情報は翻訳において有益な結果をもたらすと考えられる．

1.4 問題の分析と対応

本研究では，統計機械翻訳問題のための領域に依存した情報を獲得する問題を考える．まずはじめに，固有名詞の認識を含む形態素解析を考える．形態素解析における確率過程の有用性を確認し，ルールを用いた手法によって固有名詞の認識を行う方法を提案する．次に，統計機械翻訳におけるアラインメント問題を扱う．フレーズを生成するための手法を提案し，確率過程によってアラインメントを行う．最後に，類義語の自動抽出を行う．統計翻訳手法を用いた効果的な抽出法を提案する．

1.4.1 固有名詞の認識を含む HMM による英文形態素解析

形態素解析には確率過程モデルである**隠れマルコフモデル** (Hidden Markov Model) を用いる．HMM を用いることにより，品詞の遷移と品詞が単語を出力することを確率的に考えることが可能である．しかし，HMM だけでは固有名詞を正しく推定することが困難である．固有名詞は，1) 新規語が多い (単語が未知)，2) 領域や文脈によって意味が異なる (意味が未知)，という特徴を持つからである．

この問題を解決するために，固有名詞を出現させるルールを学習データから自動抽出し，未知語にルールを適用する方法を用いる．ルールを適用する語は，最長一致もしくは最短一致で行う．この手法により適用範囲を拡大でき，高精度に固有名詞の認識を行うことが可能であると考える．[14]

1.4.2 形態素解析とＨＭＭを用いたフレーズアラインメント

ここでは，前節までに論じた形態素解析手法を応用し，英日アラインメント問題を扱う．

表現形式が大きく異なる言語間では，単語対応を取ることが困難なため，フレーズベースの対訳法を用いる．例えば，*I was practicing tennis in the park* と**私は公園でテニスを練習していました**という 2 つの文の場合，**私は**/*I*，**練習していました**/*was practicing*，**テニスを**/*tennis*，**公園で**/*in the park* のようにフレーズ同士が対応していると考える．また，*book*_**名詞**，*book*_**動詞**のように単語に品詞を付与した形態素を用いることにより，単語の役割を限定した上でアラインメントを学習することで翻訳精度が上昇することを目指す．[15]

1.4.3 類義語の自動抽出

ニュース文章からの知識獲得として類義語の抽出を行う．一般的に類義語の抽出は手作業によって行われることが多い．類義語が単語の意味や概念を扱うものであるため，統計的な手法では抽出が困難だからである．しかし，手作業による抽出は無矛盾性を必要とし，莫大なコストがかかるという問題がある．

そこで，本研究では領域に依存した類義語の自動抽出を統計翻訳の手法を用いて行う．通常，異言語間で学習を行う統計翻訳モデルを，同一言語間で行うことにより類似関係を抽出することが可能であると考える．統計手法を用いて抽出を行うため，学習データの領域に依存した関係が得られることが期待できる．[16]

1.5 論文の構成

本研究では，以上の問題について以下の構成で論じる．第 2 章では，固有名詞の認識を含む HMM による英文形態素解析について論じる．第 3 章では，形態素解析を用いた機械翻訳におけるフレーズアラインメントを行う．第 4 章では，統計翻訳手法を用いて類義語の自動抽出を行う手法を提案する．第 5 章では，結論とする．

1.6 発表論文

・査読論文

1. 揚石 亮平, 三浦 孝夫:“Named Entity Recognition Based On A Hidden Markov Model in Part-Of-Speech Tagging”, *International Conference on the Applications of Digital Information and Web Technologies* (*ICADIWT*), 2008.
   本稿では，英文形態素解析に広く用いられてきた隠れマルコフモデル（HMM）の有用性を検証し確認する．さらに，固有名詞でありながら学習データに出現している一般語（既知語）を認識する手法として規則に基づく手法を提案する．

・研究論文

1. 揚石 亮平, 三浦 孝夫:“固有名詞の認識を含む HMM による英文形態素解析”, データ工学ワークショップ (*DEWS*), 2008.
   本稿では，英文形態素解析に広く用いられてきた隠れマルコフモデル（HMM）の有用性を検証し確認する．その後，この手法において問題となっている学習データに出現していない語（未知語）の中で，固有名詞に着目し，これを認識する手法として 2-gram を用いた手法を提案する．

2. 揚石 亮平, 三浦 孝夫:“形態素解析とHMMを用いたフレーズアラインメント”, データ工学と情報マネジメントに関するフォーラム (*DEIM*),2009.
   本研究では，統計機械翻訳の問題の１つとして知られているフレーズアラインメント問題を論じる．フレーズアラインメントでは，異なる言語表現間のフレーズを対応させることによって翻訳精度を向上させる．しかし，日本語と英語では表現形式が大きく異なり，フレーズ間の対応を取ることは容易ではない．ここでは，

隠れマルコフモデルと形態素解析を用いた対応つけを自動的に行い，その有用性を検証する．

3. 揚石 亮平, 三浦 孝夫:“統計翻訳手法を用いた類義語の自動抽出”，データ工学と情報マネジメントに関するフォーラム (*DEIM*),2010.
本研究では，類義語を自動的に抽出する方法について論じる．一般的に，類義語あるいは関連語を抽出することは容易なことではなく，人手によるコストを避けることは難しい．本研究では，複数の新聞コーパスから対応文を生成し，統計翻訳の手法により自動的に類義語を抽出する方法を提案し，実験により本手法の有用性を確認する．

第2章　固有名詞認識を含むHMMによる英文形態素解析

本研究では、形態素解析における固有名詞の認識を行うために規則ベースのテクニックを確率過程に結合する方法を提案する．本稿では実験的に英文にタグ付けをするための隠れマルコフモデル (HMM) について議論し、固有名詞の認識に焦点をあてる．規則抽出に対する n 語の連続語のルールに基づいた手法を提案し、実験結果から本手法のの有用性を示す.

2.1　まえがき

近年インターネットの急速な普及により, タグやコマンドなどの構造的手がかりを有さないテキストデータを検索解析する必要が増えている. テキストマイニングに対する近年の活発な研究では, アンケート等の自由記述データや営業日報・会議録等の，多種多様で膨大なテキストデータをどのように処理するかという問題を扱う．これまで英語文章の形態素解析のために，個々に生じる単語に品詞タグを識別する，タグ付け (tagging) 手法が論じられてきた.

多くの単語は，複数の品詞の働きを有し辞書を検査しただけでは特定できない．このため，前後の関連や共起語などから頻度解析や確率推定が効果的であるとされる．例えば，”Let me book the book of Harry Potter”という文において，単語 book は，動詞と名詞の両方の役割を持つ.

形態素解析を高精度に行うことは，(英語に限らず) 自然言語処理の重要な基本技術の 1 つである．構文や意味を正しく解析し, その意図を正しく認識するためには不可欠であると言ってよい．特に，個々のオブジェクト，振る舞い，状況を特定できる固有名詞 (Named Entity) は重要である．例えば，"東京"は都市の名前を，"カラオケ" は声なしの伴奏で歌を歌うことを，"KY(空気が読めない)"は周囲の雰囲気を察知できないことを意味している.

本研究では，英文に対して形態素解析を行う際に単純マルコフモデルに基づく確率過程モデルによる手法を適用する．確率モデルのために与えられた学習データから，動詞の book と名詞の book を正確に区別し，品詞 (Part Of Speech，POS) タグを推定する方法を学習する．我々はこれらの語を辞書から調べることが可能である．しかし，すべての固有名詞を辞書あるいは学習データから調べることは不可能である．本研究では，

辞書あるいは学習データに出現する語を既知語，そうでない語を未知語と定義する．また，各単語はそれ自身で意味を持つが，"New York"や"take into consideration"のように複数の語が 1 つの意味を持つ複合語も存在する．これら未知語に対しては，有効な推定手法が存在しない．これらの語は学習データに出現しないため頻度 0 であり，確率値 0.0 と推定される．

しかし，固有名詞を正しく認識することは，情報抽出や機械翻訳においてキーステップとなることがある．例えば，"*I like White House*"という表現からは特定の政治信条を読み取ることができるかもしれない．すなわち，文"*I like White House*"は*White House*の解釈に依存して異なる意味を持つ．また別の文"I have White House"という表現からは特定の政治信条を読み取ることができない．この事実から，本研究では既知語を固有名詞かどうかを判定するために，確率手法ではなくルールによる手法を提案する．

本研究では，形態素解析の際に固有名詞の認識を行う方法として，確率過程モデルによる手法にルールベースの技術を結合する手法を提案する．ルールの自動的抽出法について述べ，実験結果から本手法が有用であることを示す．

2.2 形態素解析

文書情報の大半は文章や図表などであり，文章は語の並びとして構成される．英語では (空白などの) 特殊文字で区切られた文字列を単語と呼ぶが，複合語 (New York 等のように複数の単語からなる語) や共起性の強い語 (連語や慣用句) 等を考慮するかどうかは，分析結果に大きな影響を与える。日本語では形態素を基本とする． 形態素とは，これ以上に細かくすると意味を失う最小の文字列 (単語) 情報を言う． 自然言語の文章に対して，各文の単語識別 (トークン化)，その語形変化 (語幹抽出や語尾変化) 品詞同定処理を総称して 形態素解析と言う． 文は形態素の並びとして構成される． 各形態素と各々が独自に有する意味を用いて，構文解析や意味解析，文脈理解などの自然言語処理に用いる．

形態素解析処理では，隣り合った形態素間の結合に関する規則を含む形態素辞書と，形態素に関する文法の知識を用いて，文を単語単位に分かち書きし，それぞれの構文上の役割を決定する． 形態素解析は，構文解析，意味解析，文脈理解などともに自然言語処理の基礎となる要素技術である．

形態素解析には多くの手法が存在する [1]． **ルールベースアプローチ**では，語・品詞とその前後に生じる語・品詞とのパターンを手作業で検出し，これを規則 (連接表) として検出する。連接表とは隣り合う形態素の結合に関する規則を列挙したものである。連接表には,"定冠詞のあとに動詞は生じない" などの規則からなる。うまく作られた連接表を利用したとき，確率過程アプローチに匹敵することが知られている．しかし，規則の生成には全域的な無矛盾性を必要とし，自明な作業ではない。必要な規則を抽出できるが，手作業による規則抽出のため，手間がかかり正確性に欠けるという欠点がある。

確率過程アプローチでは統計的手法により学習データを形態素解析して頻度を調べ，隠れマルコフモデルなどにより形態素同士の結びつきからなる状態の並びを推定する．学習データの分布情報から特徴量を抽出するため処理の汎用性が期待できる． 反面，前提となる確率過程モデルが学習データと対応するとは限らず，適用範囲が限られる可能性がある。

現在，固有名詞の認識には，外部辞書を用いたものが多く用いられている．Ji and Grishman[7] らでは，隠れマルコフモデルを用いてタグ付け後，4 つの Re-Rnaker を順次使用することによって，固有名詞を認識している．これらの Re-Rnaker とはそれぞれ構造ベース，関係ベース，イベントベース，共参照ベースでのものあり，それぞれ異なるデータを基に構成されている．この手法により，約 4 %ほどの精度が上昇している．しかし，このような手法は，外部に辞書やルールを持つため構造が複雑であり，またその構成には，大量の学習データが必要となっている．

また [8] では，各単語から語彙特徴を抽出する SVM-CMM に，LEX というトークンが固有名詞の一部であるかないかを判定するアルゴリズムを組み合わせる手法を提案している．この手法は，高い適合率と F 値を提供している．しかし，この手法も，我々のアプローチがルールベースであるという点で異なっている．

2.3 隠れマルコフモデル

2.3.1 隠れマルコフモデル

隠れマルコフモデル (Hidden Markov Model，HMM) とは，確率的な状態遷移と記号出力を備えたオートマトンである． 次の 5 項組 $M = (X, Y, A, B, \pi)$ で定義される．

1．出力記号系列 $X = \{x_1, ..., x_n\}$ であり，観測可能である．

2．状態遷移系列 $Y = \{y_1, ..., y_n\}$ であり，観測不可能である．

3．状態遷移確率分布 $A = \{a_{ij}\}$ であり，状態 y_i から状態 y_j への遷移確率である．単純マルコフを仮定している．

4．記号出力確率分布 $B = \{b_i(x_t)\}$ であり，状態 y_i で記号 x_t を出力する確率である．出力は現在の状態にのみ依存すると仮定している．

5．初期状態確率分布 $\pi = \{\pi_i\}$ であり，状態 y_i が初期状態である確率である．

すなわち，図 2.1 のようなモデルが構成される．

本稿では，あらかじめ正解が与えられた学習データからモデルを生成し，これを基に，最適な状態遷移系列の推定を行う教師あり機械学習の手法を用いることとする．

また，その推定には，Viterbi アルゴリズムを用いる．Viterbi アルゴリズムは以下のようなステップを持つ．

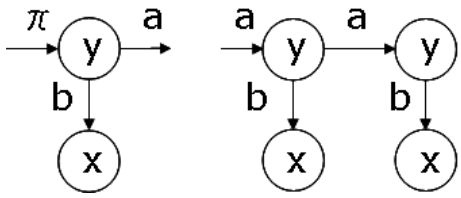

図 2.1: 隠れマルコフモデル

1．各状態 i=1,…,N に対して，変数の初期化を行う．

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(o_1) \\ \psi_1(i) &= 0 \end{aligned} \tag{2.1}$$

2．各状態の遷移ステップ t=1,…,T-1，各状態 j=1,…,N について，再帰的に計算を行う．

$$\begin{aligned} \delta_{t+1}(j) &= max_i[\delta_t(i)a_{ij}]b_j(o_{t+1}) \\ \psi_{t+1}(j) &= argmax_i[\delta_t(i)a_{ij}] \end{aligned} \tag{2.2}$$

3．再起計算の終了

$$\begin{aligned} \hat{P} &= max\delta_T(i) \\ \hat{q}_T &= argmax\delta_T(i) \end{aligned} \tag{2.3}$$

4．バックトラックによる最適状態遷移系列を復元する．t=T-1,...,1 に対して行う．

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}) \tag{2.4}$$

Viterbi アルゴリズムにより，確率値を最大にする状態遷移系列を得ることができる．

2.3.2 HMM とタグ付けの対応

英語の文章は，観測できない状態遷移系列として，品詞を持っている．例えば，*The man saw a girl* という文章の場合，その観測できない (隠れ) 状態列として，図 2.2 のように，品詞を持っていると考えられる．

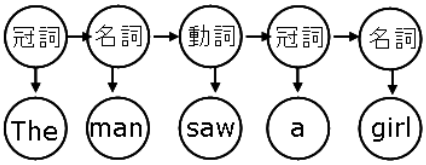

図 2.2: HMM の対応図

したがって，本稿で扱う品詞推定は，隠れマルコフモデルにおいて，状態遷移系列を品詞，出力記号系列を単語とし，学習データから，(A,B,π) を以下で示した式のように学習することにより，モデルを生成する． ただし，$P(y|x)$ は条件付き確率であり，条件 x の下で y が生起する確率を示したものである．また $C(x)$ は生起頻度であり，語 x が当該文書で出現する頻度を表す．$BOS(Beginning\ Of\ Stream)$ は初期状態であり，文の先頭であることを表す．

$$\begin{aligned} \pi_i = P(y_i|BOS) &= \frac{C(y_i)}{C(BOS)} \\ a_{ij} = P(y_j|y_i) &= \frac{C(y_i, y_j)}{C(y_i)} \\ b_i(x_t) = P(x_t|y_i) &= \frac{C(y_i, x_t)}{C(y_i)} \end{aligned} \tag{2.5}$$

2.3.3 ゼロ確率の回避

前項のモデルでは，学習データに出現しない状態遷移や記号出力のパターンは決して生じないとして，品詞の推定を行ってしまう．これでは，確率値がゼロとなるものが発生してしまい，品詞の推定が行われないおそれがある．

これを防ぐため，スムース化係数 s を導入し，スムージングを行うことにする． すなわち，(A,B,π) を以下のように再定義することである． ただし，N_T は品詞数であり，N_W は単語数である．

$$\begin{aligned} \pi_i = P(y_i|BOS) &= s \times \frac{1}{N_T} + (1-s) \times \frac{C(y_i)}{C(BOS)} \\ a_{ij} = P(y_j|y_i) &= s \times \frac{1}{N_T} + (1-s) \times \frac{C(y_i, y_j)}{C(y_i)} \\ b_i(x_t) = P(x_t|y_i) &= s \times \frac{1}{N_W} + (1-s) \times \frac{C(y_i, x_t)}{C(y_i)} \end{aligned} \tag{2.6}$$

このスムージングされた確率値を用いて，品詞の推定を行う． ただし，未知語に対する記号出力確率は一定で与えられているものとする．

2.4 固有名詞の認識

2.4.1 未知語

前章で，隠れマルコフモデルを英文形態素解析に適用する手法を述べた．しかしながら，このモデルを用いても未知語中の固有名詞を認識することは難しい．ここで，未知語とは，語自体が未知であり辞書には含まれない場合と，語は既知であるが未知の意味を含む場合に分けられる．前者の例としては *Yasuo Fukuda* のような人名が，後者の場合は *White House* や *Red Socks* などがある．後者の場合，さらに語の分かち書きに応じて異なる意味を表すことがある．例えば *New York* は地名であるが *New York Times* は新聞名である．

未知語は学習データに観測シンボルとして出現しないかもしれないので，確率過程モデルによる推定は困難であることから，我々には別の手法が必要である．すでに述べたように，未知語は名詞 (例"東京")，動詞 ("カラオケ")，形容詞 ("KY") などを含んでいる．

本研究では，Brown コーパス [5] を検査した結果，合計で 43036 語の単語を得た．そのうち 1554 語の未知語を含み，それらの 67.05%(1042) が固有名詞であった．よって，本研究では実験的に未知語は固有名詞であると仮定する．以降，既知語に着目して考える．

2.4.2 既知語として出現する固有名詞

既知語から固有名詞を推定するのが困難な理由は 2 つある．1 つは単語の**多義性**による．"book"が動詞と名詞の意味を持つように，1 つの語は複数の意味を持つ．しかし，我々は HMM によってこの役割を区別している．2 つ目の理由は固有名詞特有のものである．例えば，文"He likes Red Socks"において，我々は彼が野球チームとファッションのどちらが好きかを判断できない．一般的に，文脈解析 [10] を用いて判断を行うが，文脈解析は複雑で時間がかかるため固有名詞認識手法に適用するのは困難である．

ここで，固有名詞を含むいくつかの例を紹介する．

(1) He likes Red Socks.
(2) He has Red Socks.
(3) He likes Red Socks in Boston.
(4) She saw White Bush.
(5) She had White Bush.

(1) では，彼は野球チームかファッションのどちらかが好きであるが，(2) においてはおそらくファッションが好きである．これは一般に彼が野球チームを所有することがないからである．同様に，(3) においてはおそらく野球チームが好きである．なぜなら，そのチームの本拠地がボストンだからである．(4) では彼女は人か林を見ているが，(5) では白い林かホワイト氏の林かは判断できないが，林を見ている．

これらの例から我々は固有名詞をその周囲の語から推定できると考える．これはマルコフ過程による推定では困難である．実際，慣用表現は共起語を含んでいる．固有名詞は様々な種類の単語からなり，我々は固有名詞（"the Beatles"（冠詞と名詞:歌手），"Liberty Bell"（2つの名詞:　自由の鐘），"Deep Blue"（2つの形容詞:　スーパーコンピュータ））の内部構造まで推定できない．

2.4.3　ルールによる固有名詞の認識

固有名詞認識のためのルールを自動的に抽出する方法を述べる．

本研究で用いたBrownコーパスは固有名詞を含む品詞タグが単語ごとに与えられている．以下にいくつかの例を示す．

(1) I like Red Socks
(2) I like Red Socks in Boston
(3) I have <NE>Red Socks</NE>
(4) You see <NE> White House </NE>
(5) You see <NE> White </NE>House
(6) You see White House

(1)はファッションに関連し，(2)でもファッションに関連しているが，(3)では野球チームについて言及している．(4)では"You"が大統領の家を見ているが，(5)では"You"はホワイトさんの家を見ており，(6)では白い家を見ている．これらの例において，シンボルの<NE>と</NE>は、これらのシンボルによって囲まれた単語が固有名詞であることを意味する．

未知語中の固有名詞の認識の発見的手法として，本稿では，固有名詞に限定した連接表を用いる．これは，学習データから，固有名詞の前の1単語αとその品詞と，後ろ2単語β_1, β_2とその品詞を取り出し，固有名詞を出現させる特定のパターン$(\alpha, \beta_1, \beta_2)$．を抽出する．

例えば，抽出ルール(like_VB，in_IN，Boston_NP)は文"like HOSEI in Boston"に対応するためHOSEIを固有名詞[1]とする同様に，抽出ルール(was_BEDZ，"," who_WPS)は文"was Ageishi，who"に対応するためAgeishiを固有名詞とする．．

本研究では，固有名詞タグを含むすべてのパターンを抽出し，ルールの候補とする．しかし，有用なルールを選択する規則は存在しない．一般に，少ないルールは効果的な探索が可能であるが，再現率を低下させ，多すぎるルールは適合率を低下させてしまう．．

本研究では，**頻度**と**正確性**の2つの指標を用いる．"頻度" Qはパターンの出現頻度であり，コーパス中の固有名詞に適用されているパターンの出現回数である．"正確性" P は相対頻度であり，　パターンの全出現回数V中Q回，パターンが固有名詞に適用されたことを示している．すなわち，$P = Q/V$である．　また本研究では

[1]Brownコーパスでは VB は動詞, NP は固有名詞, NN は名詞, IN は前置詞を意味する

$s = Q \times exp(C \times P)$ を用いる．すべてのパターンをスコア s によって順位付けし，上位 N パターンをルールとして選択する．実験では，$N = 50, 100, 200, 500$ とした．

2.4.4 固有名詞の認識

本節では抽出したルールを用いて固有名詞の認識を行う手法を述べる．本研究では，固有名詞である"New York"や"New York Times"のように N グラム語と呼ばれる，N 連続語を 1 つの語として考える．これらの語は固有名詞の候補であるため，すべてのルールが適用できるかどうかを検査する．

本研究では，候補語 W と抽出ルール $(\alpha, \beta_1, \beta_2)$ に対して，W の前 1 語を α と，W の後 2 語を β_1, β_2 と比較する．もし一致していれば候補語 W を固有名詞とみなす．例えば，抽出ルール (like_VB，in_IN，Boston_NP) に対し，文"I like Red-Socks in Boston"を HMM によって解析し，品詞タグを付与する．その後，一致を行い，"Red-Socks"を固有名詞と推定する．別の文"I like Red Socks in Boston"は一致する語がないため固有名詞は存在しないと考える．

ここで，コーパスに次の 2 つの文があると仮定する．

```
I like <NE>the Beatles </NE> hotel.
I like the <NE>Hilton </NE> hotel.
```

これらの文から 2 つのルール (like_VB，hotel_NN，"."_PD) ，(the_AT，hotel_NN，"."_PD) を抽出できる． この場合，文"I like the hotel ."に 2 番目のルールが適用し，文"I like the-Beatles hotel ."には 1 番目のルールが適用する． しかし,"the-Beatles"は固有名詞であるが "the"は固有名詞ではない．この問題は最長一致問題として知られている．本研究では最長一致が正しいかどうかも実験する．

以下に本手法の手順を述べる．始めに HMM 推定によってテスト文に対し品詞タグの付与を行う．その後すべての N グラム語を生成し，長さ N の各単語に対してすべてのルールを適用する．もしルールが適用できるなら，その N グラム語の品詞を固有名詞とする．残りの語に対し，$N - 1$ として同様の方法を適用し，$N = 1$ となるまで繰り返す．

2.5 実験

本研究では，2 つの実験を行う．実験 1 では，隠れマルコフモデルを用いた品詞の推定を実験 2 では，抽出ルールによる固有名詞の認識を行う．

2.5.1 実験準備

本研究では，学習データとして Brown コーパスを用いる．Brown コーパスとは，1960 年代にアメリカ人によって書かれた英語を，約 100 万語集めて編集されたコーパスで

あり，アメリカで出版された本，新聞などから 15 のカテゴリー，500 のテキストで構成されている．

学習には，Brown コーパス全 57,082 行のうち 55,082 行 (総単語数約 110 万語) を用いる．品詞 (状態) 数は Brown コーパスに従い，その数は 311 語である．また，テストデータには，Brown コーパスの残りの 2,000 行を品詞を隠して使用する．

実験 1 では，テストデータ (Brown コーパス 2,000 行) を隠れマルコフモデルにより品詞の推定を行う．そして，その結果と，隠してある品詞を復元したものとの一致率で評価する．全単語中推定した単語の品詞が正解している割合を一致率 c として定義する．実験は学習量を変化した場合と，s 値を変化した場合の 2 つを行い，一致率 c で評価する．

実験 2 では，GoTagger という，Brill Tagger[2]を WindowsOS 上で実行できるようにしたタグ付け機を用いて，品詞を与え，それを正解とすることにする．その結果と，提案手法を用いて未知語中の固有名詞の判定を行ったものとを比較し，その適合率，再現率を計算することにより評価する．ただし，a:提案手法が固有名詞とみなした語，b:GoTagger が固有名詞とみなした語，c:提案手法と GoTagger の両方が固有名詞とみなした語である．

$$\begin{aligned} \text{適合率}\,(Precision) &= c/a \\ \text{再現率}\,(Recall) &= c/b \end{aligned} \tag{2.7}$$

ここでは，コーパスから自動抽出したパターンを用いる．本研究では，31318 件のパターンを得た．実験では，スコア s により上位 N をルールとして選択する．$N = 50, 100, 150, 200, 500$ と変化させて実験を行う．以下に上位 10 件の抽出ルールを示す．

```
(".", "," , "who\verb+_+WPS"), (",", "said_VBD", ".")
(",", ",", "Jr._NP"), ("by_IN", ","", "who_WPS")
(",",",", "and_CC"), ("(", ",)", "of_IN")
("the_AT, "&_CC", "Sharpe_NP"),
("Part_NN", "of_IN", "the_AT)
("in_IN", ",", "1960_CD"), ("in_IN", "Island_NN", ".")
```

その後，抽出ルールの適用範囲を最大 5 連続語までとし，最長一致もしくは最短一致を適用する．ここでは適合率，再現率とともに F 値[3]を用いて評価を行う．実験では，HMM のみの場合 (未知語に対する仮定，ルールを用いない) と HMM に未知語の仮定を追加した場合と HMM にすべての手法を適用した場合の 3 種類の実験を行う．最長一致，最短一致に関しても実験を行う．

また，ルール精度 k_1/k を用いる．各ルールに対し，k 件のルール適用回数中 k_1 の正解件数を計算する．

[2]Eric Brill が開発したタグ付け機であり，Brown コーパスを基に製作した品詞辞書，文法辞書などによりタグ付けを行う．

[3]F 値 $= 2 \times Precision \times Recall/(Precision + Recall)$

2.5.2 実験結果

実験 1 の，学習量を変化した場合と，s 値を変化した場合の一致率の変化を (表 2.1, 表 2.2) に示す.

学習量を変化させた場合，全てを学習した場合が最も一致率が高く一致率=90.81%であった．また，学習量を 55,082 行として s 値を変化させた場合，s=0.01 %で最も一致率が高く一致率=93.41%であった.

学習量 (行)	一致率 (%)
1000	61.79
5000	83.12
10000	87.11
20000	89.19
30000	90.02
40000	90.46
45000	90.61
50000	90.67
52000	90.69
55082	90.81

表 2.1: 学習量変化

s 値 (%)	一致率 (%)
10	92.31
5	92.88
1	93.27
0.5	93.34
0.1	93.40
0.05	93.40
0.01	93.41
0.005	93.41
0.001	93.40
0.0001	93.37

表 2.2: スムース化係数 s 変化

固有名詞の認識における適合率を再現率を表 2.3 に示す．HMM のみの場合，推定出来ない語はスキップする．HMM に未知語の仮定を適用する場合は未知語としてタグ付けする．すべての行中の n 連続語に対して最長一致，最短一致を適用した結果を示す.

実験から，本研究では，代名詞 ("he", "you")，冠詞 ("a", "an", "the"，前置詞 ("as", "in", "with")，接続詞 ("but, and") BE 動詞 ("is", "were") はほとんど固有名詞に含まれないことがわかった．実際，これらの語は学習データ中の固有名詞 39138 語中 45 語 (0.11%) しか存在していなかった.

よって本研究では，これらの単語を含む語を n 連続語から取り除く**フィルタリング**を行う.

フィルタリングなしの場合，適合率と F 値が最も高いのは最短一致でルール数を 50 としたときであり，再現率が高いのは最長一致でルール数を 500 としたときであった．このことは，抽出ルールのランク付けにより正しいルールを抽出可能だが，適用範囲に対しては有用でないことを示している．同様に，フィルタリングを用いた場合では最も高い適合率はルール数を 50 としたとき，最も高い F 値はルール数を 200 としたときであり，最長一致かどうかは影響しなかった (表 2.4)．表 2.3 と比較して，フィルタリングを用いることによりルールの精度は上昇している．.

実験	適合率	再現率	F 値	ルール精度
HMM	95.34	75.80	84.45	
+Unknown	84.88	88.57	86.69	
(最長一致)				
50 rules	84.04	89.19	86.54	69.41
100 rules	83.73	89.37	86.45	67.87
200 rules	83.52	89.55	86.43	69.89
500 rules	81.52	90.30	85.69	58.48
(最短一致)				
50 rules	84.06	89.21	86.56	69.82
100 rules	83.80	89.46	86.54	68.98
200 rules	83.60	89.66	86.52	70.80
500 rules	81.67	90.51	85.86	59.28

表 2.3: 適合率と再現率 (フィルタリングなし)

2.5.3 考察

実験 1 の学習量の変化においては，すべてを学習した場合が最も一致率が高くなっている．これは，学習量を増やすことにより未知語の数を減少させることができたからであることが考えられる．未知語が多く存在するとそれに比例してゼロ確率が発生してしまう．それにより推定が行われない可能性が高くなってしまうからであると考えられる．また，実験 1 の s 値の変化においては，s 値を変化させることにより 2.6%の利得を得ることができている．未知語の出力確率を一定で与えているため，その推定は状態遷移確率に依存している．つまり，未知語に対しては，最も頻繁に現れるパターンの品詞を選択することができたため，一致率の上昇につながったと考えられる．

学習量を 55,082 行，s 値を 0.01%とした場合の不正解総数は 2863 語であり，その中に未知語が半分 (1451 語) 含まれている．また，未知語総数は 2188 語であったので，約 66 %が誤りとなっている．このことから，隠れマルコフモデルを用いて品詞の推定を行う場合，未知語が多く存在することがその結果に悪影響を与えると考えられる．未知語を生じないように学習させた場合の一致率が 97.35%だったことからもこの推測が正しいといえよう．したがって，隠れマルコフモデルをタグ付けに用いる場合，学習データの語分布に依存するということになる．

実験 2 の固有名詞認識では，ルールベース手法を適用することにより推定精度が上昇している．特にルール自動抽出法はすべてのパターンの探索を避けるため効果的な手法である．

1 つの問題として，本研究では多くの連続語を生成し，ルールの信頼性を改善すべきである．瀕死フィルタリングによって候補数を減らすことが可能である．また，ルールの信頼性を向上させることが可能である．実験結果から信頼性の高い方法でルール

実験	適合率	再現率	F 値	ルール精度
(最長一致)				
50	84.54	88.98	86.70	84.29
100	84.49	89.16	86.76	85.31
200	84.51	89.30	86.84	88.70
500	83.74	89.80	86.67	79.68
(最短一致)				
50	84.54	88.98	86.70	84.29
100	84.49	89.16	86.76	85.31
200	84.51	89.30	86.84	88.26
500	83.76	89.78	86.65	79.62

表 2.4: 適合率と再現率 (フィルタリングあり)

を用いることにより固有名詞の認識が可能であることがわかる.

2.6 結論

本研究では，固有名詞の認識手法を提案した．学習データから HMM モデルを構成し，固有名詞認識のためのルールの自動抽出を行った．その後 HMM によって品詞を推定したテストデータに対してルールを適用した．抽出したルール集合を改良するため，品詞フィルタリングの手法を提案し，実験から手法が有用であることを確認した.

第3章　形態素解析とHMMを用いたフレーズアラインメント

本稿では，統計機械翻訳の問題の１つとして知られているフレーズアラインメント問題を論じる．フレーズアラインメントでは，異なる言語表現間のフレーズを対応させることによって翻訳精度を向上させる．しかし，日本語と英語では表現形式が大きく異なり，フレーズ間の対応を取ることは容易ではない．ここでは，隠れマルコフモデルと形態素解析を用いた対応づけを自動的に行い，その有用性を検証する．

3.1　まえがき

近年，インターネットの普及により，他言語による情報に接することが容易になってきた．それに伴い，機械翻訳に注目が集まるようになり，様々な研究が行われ，その学習目的のための大規模対訳コーパスが利用可能となった．そして，現在では，対訳コーパスからの学習方法として，統計翻訳 [2](Statistical Machine Translation, SMT) が著しい発展を遂げている．

SMT の１つの問題として，アラインメント問題が挙げられる。ここで，アラインメントとは，ある単語とその語と等価な意味を持つ翻訳語との対応を取ることである．例えば，*I like apples* と**私はりんごが好き**という２つの文の単語アラインメントを考えた場合，*I*(**私は**)，*like*(**好き**)，*apples*(**りんごが**) となる．そして，アラインメントの精度が上昇することにより，翻訳精度が上昇すると考えられる．

しかし，アラインメントを取ることは容易なことではない．例えば，辞書を用いた場合，*book* という単語が**本**，**予約する**など複数の意味を持つように，多くの英単語は一義的に意味を決定することができない．これは，形態素解析の問題と関連している．

また，たとえ形態素解析を行ったとしても，日本語と英語のような表現形式が大きく異なる言語では，単語間の対応を取ることは容易ではない．例えば，*I was practicing tennis in the park* と**私は公園でテニスを練習していました**という２つの文の場合，英単語 *was* に対応する日本語は**た**であるが，これは**練習する**という語の助詞として扱われてしまっている．また，英単語 *in*，*the* に対応するような日本語は存在していない．したがって，本研究では，意味を持つ単位としてフレーズを用い，フレーズをベースとしたアラインメントを考える．すなわち，２つの対訳文を *I/was practicing/tennis/in the park* と**私は/公園で/テニスを/練習していました**というようにフレーズに分割し，

その後，対応付けを行う．

本研究では，隠れマルコフモデルと形態素解析を用いた手法により，英語と日本語のフレーズアラインメントを獲得する手法を提案する．その際，形態素解析と辞書を用いた発見的方法により，フレーズを生成し，フレーズ間の対応付け行う．そして，提案手法が効果的であることを実験で検証する．

第 2 章で，フレーズアラインメントのために用いた隠れマルコフモデルの適用方法を示す．第 3 章では，形態素解析と辞書によるフレーズの対応付け方法を述べる．第 4 章では，実験結果を述べ，本手法の有効性を示す．

3.2 隠れマルコフモデル

3.2.1 隠れマルコフモデル

隠れマルコフモデル (Hidden Markov Model，HMM)[19] とは，確率的な状態遷移と記号出力を備えたオートマトンである． 次の 5 項組 $M = (X, Y, A, B, \pi)$ で定義される．

1．出力記号系列 $X = \{x_1, ..., x_n\}$ であり，観測可能である．

2．状態遷移系列 $Y = \{y_1, ..., y_n\}$ であり，観測不可能である．

3．状態遷移確率分布 $A = \{a_{ij}\}$ であり，状態 y_i から状態 y_j への遷移確率である．単純マルコフを仮定している．

4．記号出力確率分布 $B = \{b_i(x_t)\}$ であり，状態 y_i で記号 x_t を出力する確率である．出力は現在の状態にのみ依存すると仮定している．

5．初期状態確率分布 $\pi = \{\pi_i\}$ であり，状態 y_i が初期状態である確率である．

すなわち，図 3.1 のようなモデルが構成される．

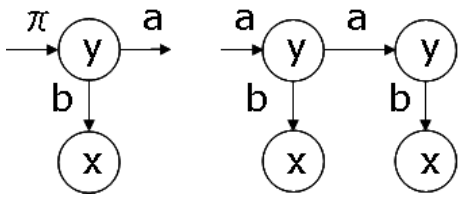

図 3.1: 隠れマルコフモデル

本稿では，対訳コーパスからモデルを生成し，これを基に，最適な状態遷移系列の推定を行う教師あり機械学習の手法を用いることとする．

また，その推定には，Viterbi アルゴリズムを用いる．Viterbi アルゴリズムは以下のようなステップを持つ．

1．各状態 i=1,⋯,N に対して，変数の初期化を行う．

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(o_1) \\ \psi_1(i) &= 0 \end{aligned} \tag{3.1}$$

2．各状態の遷移ステップ t=1,⋯,T-1，各状態 j=1,⋯,N について，再帰的に計算を行う．

$$\begin{aligned} \delta_{t+1}(j) &= max_i[\delta_t(i)a_{ij}]b_j(o_{t+1}) \\ \psi_{t+1}(j) &= argmax_i[\delta_t(i)a_{ij}] \end{aligned} \tag{3.2}$$

3．再起計算の終了

$$\begin{aligned} \hat{P} &= max\delta_T(i) \\ \hat{q}_T &= argmax\delta_T(i) \end{aligned} \tag{3.3}$$

4．バックトラックによる最適状態遷移系列を復元する．t=T-1,...,1 に対して行う．

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}) \tag{3.4}$$

Viterbi アルゴリズムにより，確率値を最大にする状態遷移系列を得ることができる．

3.2.2 HMM フレーズアラインメント

この節では，フレーズアラインメントを HMM に対応させる方法を述べる．フレーズ化の方法については，次章で説明する．

英語文は，観測できない状態遷移系列として，日本語文を持っていると考える．例えば，*I was practicing tennis in the park* という文章の場合，その観測できない (隠れ) 状態列として，**私は公園でテニスを練習していました**を持っていると考えられる．そして，これらの文章をそれぞれフレーズに分割し，フレーズ間の対応を取ると，図 3.2 のようになる．

この場合，状態遷移として，**私は**，**練習していました**，**テニスを**，**公園で**の順に状態が遷移していったと考え，その各状態での記号出力として，**私は**/*I*，**練習していました**/*was practicing*，**テニスを**/*tennis*，**公園で**/*in the park* をそれぞれの状態が出力したと考える．

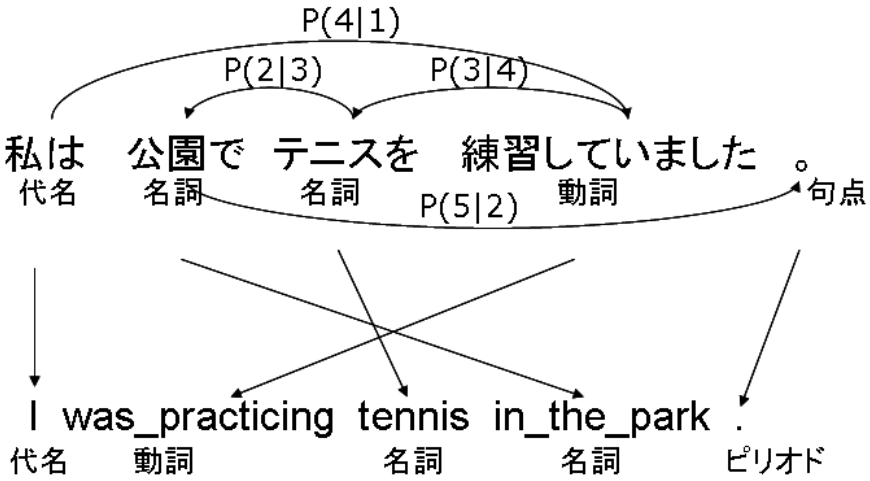

図 3.2: HMM の対応図

したがって，本稿で扱う品詞推定は，隠れマルコフモデルにおいて，状態遷移系列を日本語フレーズ，出力記号系列を英語フレーズとし，学習データから，(A,B,π) を以下で示した式のように学習することにより，モデルを生成する． ただし，$P(y|x)$ は条件付き確率であり，条件 x の下で y が生起する確率を示したものである．また $C(x)$ は生起頻度であり，語 x が当該文書で出現する頻度を表す．$BOS(Beginning\ Of\ Stream)$ は初期状態であり，文の先頭であることを表す．

$$\begin{aligned} \pi_i &= P(y_i|BOS) = \frac{C(y_i)}{C(BOS)} \\ a_{ij} &= P(y_j|y_i) = \frac{C(y_i, y_j)}{C(y_i)} \\ b_i(x_t) &= P(x_t|y_i) = \frac{C(y_i, x_t)}{C(y_i)} \end{aligned} \tag{3.5}$$

3.3 HMM のためのフレージング

この章では，前章に述べた HMM フレーズアラインメントモデルを適用するために，対訳コーパスを，フレーズに分割し，対応付けを行う手法を述べる．ここで，HMM は観測列 (英語文) と隠れ状態列 (日本語文) のそれぞれの語の対応が明確でないと学習できないことに注意されたい．

3.3.1 フレーズ生成

まず，対訳コーパスをそれぞれ形態素解析する．英文形態素解析には，揚石ら [14] の手法を，日本語形態素解析には， MeCab[12] を用いる．

揚石らの手法では，HMM によって英文に品詞 (タグ) を付与した後，自動的に抽出したルールにより固有名詞の再推定を行っている．固有名詞を正しく認識できることは，翻訳にとって重要であると考えられる．例えば，*While House* が，**白い家**か**ホワイトハウス**なのかでは大きく意味が異なる．したがって，本稿ではこの手法を採用する．

この手法により，*I was practicing tennis in the park* という文の形態素解析を行うと，*I_PPSS was_BEDZ practicing_VBG tennis_NN in_IN the_AT park_NN* となる．ここで，*PPSS* は一人称代名詞を，*BEDZ* は be 動詞過去形を，*VBG* は動名詞を，*IN* は前置詞を，*AT* は冠詞を，*NN* は名詞を意味している．

このようにして得られた形態素をもとに，発見的規則によりフレーズ化を行う．例えば，冠詞 (*the*) に続く名詞 (*park*) は，フレーズ (*the_park*) とし，その品詞を名詞とする，といった形である．

また，日本語形態素解析に用いた Mecab は，京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンであり，CRF によってパラメータの推定を行っている．

MeCab により，**私は公園でテニスを練習していました**という文の形態素解析を行うと，表 3.1 のような結果となる．これは，左から順に，表層形，読み，原型，品詞，細品詞 1，細品詞 2，となっている．

《私》《ワタシ》《私》《名詞》《一般》《代名詞》
《は》《ハ》《は》《助詞》《*》《係助詞》
《公園》《コウエン》《公園》《名詞》《*》《一般》
《で》《デ》《で》《助詞》《一般》《格助詞》
《テニス》《テニス》《テニス》《名詞》《*》《一般》
《を》《ヲ》《を》《助詞》《一般》《格助詞》
《練習》《レンシュウ》《練習》《名詞》《*》《サ変接続》
《し》《シ》《する》《動詞》《*》《自立》
《て》《テ》《て》《助詞》《*》《接続助詞》
《い》《イ》《いる》《動詞》《*》《非自立》
《まし》《マシ》《ます》《助動詞》《*》《*》
《た》《タ》《た》《助動詞》《*》《*》
《。》《。》《。》《記号》《*》《句点》

表 3.1: MeCab による形態素解析

この結果をもとに，英語文の場合と同様にして，形態素によるフレーズ化を行う．そ

して，これらの英語と日本語のフレーズ化により，図 3.2 のようなフレーズを獲得することができる．

3.3.2 フレーズの対応付け

次に，以下の得られたフレーズ間の対応付けを行う方法について説明する．

- *I*/**代名詞** *was_practicing*/**動詞** *tennis*/**名詞** *in_the_park*/**名詞**
- **私は/代名詞 公園で/名詞 テニスを/名詞 練習していました/動詞**

まず，品詞によってユニークに決定できる場合，それらの語を対応付ける．すなわち，*I* と**私は**が，*was_practicing* と**練習していました**が対応していると考える．

次に，品詞によって決定することができなかった残りの候補を辞書を引くことによって対応付けをする．ここで，辞書には英辞郎 [18] を用いた．英辞郎の構成は表 3.2 のようになっている．

park 他動-1 : 〜を駐車 {ちゅうしゃ} させる、駐車 {ちゅうしゃ} する
park 他動-2 : 〜を置いておく
park 名-1 : 公園 {こうえん} ■・Karl took a walk in the park after lunch. カールは昼食の後、公園を散歩した。
park 名-2 : 遊園地 {ゆうえんち}
park 名-3 : 大庭園 {だい ていえん}
park 名-4 : 競技場 {きょうぎじょう}
park 名-5 : 駐車場 {ちゅうしゃじょう}
park 名-6 : 〈米話〉野球場 {やきゅう じょう}
park 自動-1 : 駐車 {ちゅうしゃ} する
park 自動-2 : 《野球》(ホームランを) 打つ
park : 【レベル】1、【発音】p ɑ ':(r)k、【@】パーク、 【変化】《動》parks — parking — parked
park a bicycle : (自転車 {じてんしゃ} を) 駐輪 {ちゅうりん} する
park a car in a garage : 駐車場 {ちゅうしゃじょう} に車を止める
park a car in a narrow space : 狭い場所 {ばしょ} に駐車 {ちゅうしゃ} する

表 3.2: 英辞郎の例

ここで，{}は品詞を，: より後が翻訳語を表している．また，■より後ろは用例を表している．

本研究では，: と■の間の語に着目し，品詞が一致し，かつ翻訳語が一致するフレーズを対応付けを行う．そして，最後にもう一度残った候補の中から，品詞によってユニークに決定できるフレーズを対応付ける．

これにより，対応付けされたフレーズ，**私は**/*I*，**練習していました**/*was practicing*，**テニスを**/*tennis*，**公園で**/*in the park*，を得ることができる．

3.4 実験

3.4.1 準備

本研究では，対訳コーパスとして読売新聞と The Daily Yomiuri の対訳文 [11] を用いる．これは，内山らによって作成された対訳コーパスで，1989 年から 2001 年までの読売新聞と The Daily Yomiuri からなっており、文対応ペアは 1 対 1 対応が約 15 万ペアある．このうち，無作為に選択した 2,000 行をテストデータとし，残りを学習データに当てる．対訳コーパスの例を表 3.3 に挙げる．

<J> 欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつつある。</J>
<E>Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen. </E>
<J> 我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。/<J>
<E>We recognize the importance of improved market access for economic progress in Russia./<E>

表 3.3: 対訳コーパス:読売新聞–The Daily Yomiuri

3.4.2 評価方法

Word Error Rate (WER)[13] は語順を考慮した不一致率であり，正解との編集距離によって計算される．

$$WER = \frac{\sum_i (\text{挿入語数}\, i + \text{削除語数}\, i + \text{置換語数}\, i)}{\sum_i \text{参照訳}\, i \,\text{の語数}} \tag{3.6}$$

また，Position independent word Error Rate(PER) は語順を無視し、文を単語集合とした不一致率である．これは，WER では翻訳が正しくても，正解と語順が著しく異なる場合に結果が悪くなってしまうため，語順を考慮せず，語のレベルでその翻訳が正しいかを判断するための基準である．

$$PER = 1 - \frac{\sum_i (\text{翻訳文}\, i \text{・参照訳}\, i \text{間の一致語数})}{\sum_i \text{参照訳}\, i \text{の語数}} \tag{3.7}$$

どちらも，誤り率であるので，数値が低いほど良い結果であるといえる．本研究では，語を並び替える操作は考えていないため，PER の結果が良いことが望ましい．

3.4.3 実験結果

テストデータ 2,000 行に対し，Viterbi アルゴリズムを用いて，日本語の推定を行い，WER，PER の値を測定した．実験結果を表 3.4 に示す．

	WER	PER
読売新聞	1.13	0.64

表 3.4: 実験結果

PER の値が 0.74 であるので，およそ 4 語に 1 語の割合で正しい日本語を推定できていると考えられる．また，本実験では，並び替えは考えていないため，WER の値が 1 を超えてしまっているこれは，翻訳語と正解語が違うことに加えて，翻訳語が，正解語よりも多いため，削除を多くしてしまっているためである．

3.4.4 考察

PER の値が，低い理由として，学習不足が考えられる．実際に学習データとして 148,000 行与えているが，対応付けを自動的に行い，学習に使用することができたデータ数は，1315 行であった．これにより，多くの未知語 (学習データに出現しない語) が増えてしまったため，精度が悪化していると考えられる．学習がうまくいく例としては，

- We/PPSS recognize/VB the-importance/NN of-improved/VBN market-access/NN for-economic-progress/NN in-Russia/NP ./.
- 我々は/代名詞 ロシアの/固有名詞 経済発展にとって/名詞 改善された/動詞 市場アクセス-が/名詞 重要であることを/名詞 認識する/動詞 。/句点

の対訳文が挙げられる．この例のように，英語文と日本語文のそれぞれのフレーズが，順序が異なるだけであれば学習はうまくいく．この 2 つの文の対応付けを行うと，*We/* **我々は** *recognize/* **認識する** *the-importance/* **重要であることを** *of-improved/* **改善された** *market-access/* **市場アクセスが** *for-economic-progress/* **経済発展にとって** *in-Russia/* **ロシアの** *./。* となる．しかし，

- International-terrorism/NP is/BEZ a-grave-threat/NN to-world-peace/NN and/CC security/NN ./.
- 国際テロは/名詞 世界の/名詞 平和と/名詞 安全に対する/名詞 重大な脅威-だ/名詞 。/句点

のような場合は対応付けを行うことができない．これはまず，英語文の動詞 *is* に対応する日本語が存在していない．このように英語と日本語の文法構造が異なる文では，うまく学習することができない．また，英語の *to-world-peace/NN and/CC security/NN* と日本語の**世界の/名詞 平和と/名詞 安全に対する/名詞**のフレーズ構造が異なってしまっている．これは，英文では，*to-world-peace/NN*(**世界平和**) と security/NN(安全) と解釈したが，日本語文では，**世界の** (**平和と安全**) と解釈してしまっているためである．このようにフレーズ化した際の違いによっても，学習が困難となっている．

さらに，テストデータを調査してみたところ，テスト語 32606 語のうち，19471 語が未知語と判定されていた．未知語を除いて PER の値を計算すると，PER=0.41 であったので，未知語を減らすことによって，精度が上昇すると考えられる．したがって，発見的ルールを見直し，より多くの学習データを使用することができれば，それに伴い，精度を上げられる可能性があると考えられる．

3.5 結論

本研究では，英語と日本語のフレーズアラインメントを生成する方法として，形態素解析を隠れマルコフモデルを用いた方法を提案した．実際に実験を行い，その結果から提案手法がフレーズアラインメントを生成する方法として期待できることを確認した．

第4章　統計翻訳手法を用いた類義語の自動抽出

本研究では，類義語を自動的に抽出する方法について論じる．一般的に，類義語あるいは関連語を抽出することは容易なことではなく，人手によるコストを避けることは難しい．本研究では複数のコーパスから文対応を生成し，統計翻訳の手法により自動的に類義語を抽出する方法を提案する．実験により本手法の有用性を確認する．

4.1　まえがき

近年，インターネットの普及により，誰もが容易に複数の言語による情報を得られるようになった．これらを容易に理解するために機械翻訳に注目が集まり，様々な研究が行われている．現在では大規模対訳コーパスが利用可能なことから統計機械翻訳 [2](Statistical Machine Translation，SMT) が著しい発展を遂げている．

SMT 手法のうち，Brown ら [4] による IBM Model では，英仏対訳コーパスから確率的な仏英単語辞書を語彙確率として学習する．仏文を英文に翻訳する際には，IBM Model は入力仏文に対して単語ごとにすべての語彙確率を計算し，最も確率が高い英文を出力する．

一方で類義語抽出に関する統計的研究は多く行われていない．類義語が単語の意味や概念を扱うものであり，人手によるコストを避けることが難しいからである．しかし背景知識を必要とする類義語は存在している．例えば以下のようなニュース記事の例の場合，

- **麻生太郎，総裁選に立つと表明**
- **麻生太郎氏が総裁選に出馬すると表明した**

この 2 文では**立つ**と**出馬する**の意味が類似する．特定の背景知識を仮定しないとき、例えば大辞林によれば，**立つ**と**出馬する**はそれぞれ**座ったり横になったりしていた人が足を伸ばして自分の体を垂直の姿勢にする**，**馬に乗って出かける**という意味を有し、同義語・類義語ではない．一方，これら 2 語はニュースという領域において類義語である．本研究では類義語を**一般的類義語**（背景知識がなく類義であるもの）と**領域依存型類義語**（背景知識があって類義であるもの）の 2 つに分類し，領域依存型類義語に着

目して考える．領域依存型類義語は文書分類や情報検索において重要である．例えば，領域依存型類義語辞書を情報検索に用いると，与えられた検索語の類義語から領域に依存して情報を拡大することが可能である．

本研究の目的は，領域依存型類義語の自動抽出を行うことである．統計的手法はコーパスに依存した学習で有用であり，我々は，統計翻訳手法による領域依存型類義語の自動抽出法を提案する．統計翻訳手法を用いて特定領域のコーパスから語彙確率の学習を行い，ある単語に対して語彙確率の高い語をその語の類義語と見なす．本研究では固有名詞の重要性を考慮した複数コーパスからの文対応生成方法についても提案する．揚石ら [14] は確率過程モデルによる形態素解析結果を固有名詞に関して再推定を行うことで，形態素解析精度が改善されることを示した．固有名詞が重要な役割を果たしている例である．本稿では提案手法が効果的であることを実験により検証する．

第 2 章で，関連研究について述べ，第 3 章では，ニュースコーパスの文対応生成方法を述べる．第 4 章では，統計翻訳手法を説明し，類義語の抽出方法を述べる．第 5 章では，実験結果を述べ，本手法の有効性を示す．

4.2 関連研究

類義語に関する研究としては，George ら [6] による Wordnet がある．Wordnet は英語の概念辞書であり，Synset 概念を導入し類義関係を語の組で表現する．各 Synset が 1 つの概念に対応する．Wordnet では各語が複数の Setset に属してよいため，語の上下関係を表現できる．辞書の構築は手作業で行われているため，大きな変更が起こってしまうと莫大な人手コストがかかるという問題がある．

Wordnet を日本語化したものとして，Kanzaki ら [9] による日本語 Wordnet がある．日本語 Wordnet は Wordnet と同様の構造を持つ．日本語 Wordnet の例として，検索語**立つ**を入力した結果を表 4.1 に示す．

表 4.1 中の各数字は Synset の ID を表す．表 4.1 から，Wordnet の類義語は**起きる**や**立ち去る**といった一般的類義語であることが分かる． よって Wordnet は本研究で考える領域依存型類義語とは性質が異なる．

類義語の抽出を領域依存で自動的に行った研究としては，中渡瀬 [21] の手法がある．中渡瀬の手法はコーパスから複合語を取り出し複合語の修飾関係などを考慮する．複合語を分割し，2 部グラフで表現することにより類義語を抽出する．例えば，**国際戦略**，**国際不況**，**世界不況**，**世界戦略**という 4 語がある場合，（**国際**，**世界**）と（**戦略**，**不況**）という 2 部グラフを構成する．このとき各グループ（**国際**，**世界**），（**戦略**，**不況**）が類義語であると考える．中渡瀬の手法には複合語が不可欠であるため，本研究が考える単語ごとの類義語の抽出には適さない．

01983264-v: 起きる, 立っち, 起立, 立上がる, 立つ, 起き上がる, 立上る
01848718-v: 去る, 退場, 離れさる, 出立, 発す, 走去る, 出発, 走りさる, 離れ去る, 発つ, 立ちさる, 立去る, 離去る, 立ち去る, 出立つ, 立つ, 消え失せる
02618688-v: 成り立つ, 発つ, 成立つ, 立つ
01546111-v: 立ち上がる, 起きあがる, 起上る, 起きる, 立っち, 起上がる, おっ立つ, 押っ立つ, 起立, 立上がる, 立つ, 起き上がる, 起き上る, 立上る, 起つ
02734488-v: 御座る, 御座ある, ござ在る, 位置, 御座在る, いらっしゃる, 居る, ござ有る, 御座有る, 在る, 立つ, 御座居る, ある, 有る

表 4.1: 日本語 Wordnet の例

4.3 ニュース記事の文章対応

本章では 2 つのニュースコーパスから文の対応を得る方法について論じる．次章で述べる統計翻訳手法を用いるためには，対訳コーパス (Parallel Corpus) が必要である． 本研究では日本語から日本語への翻訳モデルを学習し，得られた語彙確率を類似度として用いるため，日本語と日本語の対訳コーパスが必要である．しかし大規模な日本語間対訳コーパスは存在しないため，本研究では自動的な文対応生成方法を提案する．

4.3.1 記事の対応生成

本節では記事の対応生成方法を述べる．本研究において記事とは，タイトル，本文，日付からなるものとする．本研究で用いた朝日新聞と読売新聞の例を表 4.2，表 4.3 に示す．

表中の＼ＡＦ＼，＼ＹＦ＼が日付を，＼Ｔ１＼がタイトルを，＼Ｔ２＼が本文を表す．本研究では，対応する記事は同一の日付であり，タイトルの内容が類似すると考える．本研究では同一の日付のタイトルを取り出し，タイトルの比較を行うことで記事の対応を得る方法を用いる．

タイトルの比較を行う方法について述べる．表 4.2，表 4.3 からタイトルを抽出しそれぞれ形態素解析を行うと，

(A) 麻生 (固有名詞) 太郎 (固有名詞) 氏 (接尾) が (助詞) 総裁選 (名詞) に (助詞) 出馬 (名詞) を (助詞) 表明 (名詞)

＼ＡＦ＼	２００７０９０１	２００７０９０１
＼Ｔ１＼	麻生太郎氏が総裁選に 出馬を表明	「最強」法大か、粘りのＯＳか アメフット・ライスボウル
＼Ｔ２＼	麻生太郎氏が総裁選に 出馬すると発表した. その足で千葉駅に向かい, 麻生氏は演説を行った.・・・	アメリカンフットボール日本一を 決める第２４回日本選手権は ・・・

表 4.2: 朝日新聞

＼ＹＦ＼	２００７０９０１	２００７０９０１
＼Ｔ１＼	千葉でひき逃げ事件の 容疑者を逮捕	麻生氏総裁選に立つ
＼Ｔ２＼	３１日未明に発生した ひき逃げ事件の容疑者, として千葉県警は３８歳 男性を逮捕した. ・・・	麻生太郎氏は総裁選に 立候補することを発表した. それに伴い，麻生氏は 千葉駅前で熱弁を振った. 総裁候補が千葉で演説を 行うのは初めてのこと.・・・

表 4.3: 読売新聞

(B) 「(記号) 最強 (名詞) 」(記号) 法大 (固有名詞) か (助詞) 、(記号) 粘り (名詞) の (助詞) ＯＳ (固有名詞) か (助詞) アメフット・ライスボウル (固有名詞)

(C) 千葉 (固有名詞) で (助詞) ひき逃げ (名詞) 事件 (名詞) の (助詞) 容疑者 (名詞) を (助詞) 逮捕 (名詞)

(D) 麻生 (固有名詞) 氏 (接尾) 総裁選 (名詞) に (助詞) 立つ (動詞)

となる．形態素解析結果からタイトルには助詞を除くと名詞，固有名詞が多く使われることが分かる．したがって本研究では名詞，固有名詞の一致度によってタイトルの比較を行う．一致度 M を以下の式で定義する．

$$M_1 = \frac{\sum_i \sum_j Match(w_i^F, w_j^E)}{max(\sum_i C(w_i^F), \sum_j C(w_j^E))} \tag{4.1}$$

(4.1) 式において，w^F は朝日新聞中の名詞または固有名詞，w^E は読売新聞中の名詞または固有名詞を表す．また，$\mathtt{Match}(w_i^F, w_j^E)$ は一致した名詞または固有名詞の頻度を表す．$C(w_i)$ は生起頻度であり，語 (w_i) が該当文書で出現する頻度を表す．したがって

(4.1) 式は，名詞または固有名詞の相対一致度を表す．一致度 M を用いて (A) と (C)，(A) と (D)，(B) と (C)，(B) と (D) の比較を行うと，(A) と (C) で $M_1 = 0/6 = 0.0$，(A) と (D) で $M_1 = 2/6 = 0.33$，(B) と (C) で $M_1 = 0/5 = 0.0$，(B) と (D) で $M_1 = 0/5 = 0.0$ となる．この例が表すように，タイトル (A) を含む記事とタイトル (D) を含む記事が対応すると考える．本研究では (4.1) 式を用いて，すべての同一の日付のタイトルの比較を行い，閾値を用いることで記事の対応を得る．

4.3.2 対応記事中の本文対応生成

本節では，得られた記事の対応からの本文対応生成方法について述べる．前節に示したタイトル対応のうち，(A) と (D) の記事対応が存在すると仮定する．

表 4.2 のタイトル (A) の記事本文

(a1) 麻生太郎氏が総裁選に出馬すると発表した

(a2) その足で千葉駅に向かい，麻生氏は演説を行った

と表 4.3 のタイトル (D) の記事本文

(d1) 麻生太郎氏は総裁選に立候補することを発表した

(d2) それに伴い，麻生氏は千葉駅前で熱弁を振った

(d3) 総裁候補が千葉で演説を行うのは初めてのこと

の対応を考える．(4.1) 式を用いて対応する可能性のある文は，(a1) と (d1)($M_1 = 0.75$)，(a2) と (d2)($M_1 = 0.4$)，(a2) と (d3)($M_1 = 0.4$)) である．(4.1) 式を用いると (a2) に対して (d2) と (d3) の 2 つの候補が存在するが，(a2) が実際に対応する文は (d2) である．(a2) の対応文が (d2) であることは，**演説**という名詞よりも**千葉**や**麻生**という固有名詞が一致することから判断可能である．さらに**麻生**というより重要な語が一致する点からも判断可能である．**麻生**という語が**千葉**という語よりも重要であるということは，表 4.3 から**麻生**は 2 記事中 1 回，**千葉**は 2 記事中 2 回出現することから分かる．本研究では，固有名詞が一致することに重みをつけ，固有名詞ごとに重みを変化させるために，TF*IDF 法[1] [20] の IDF 値による重みを用いる．本研究では，固有名詞 w_i の IDF_{w_i} を以下の式のように定義する．

$$IDF_{w_i} = log\frac{D}{\{d : d \ni w_i\}} \tag{4.2}$$

[1] TF*IDF 法とは，情報検索などに用いられる文章中の重要語を抽出するための重み付け手法である．TF は文章中の単語の出現頻度を表し，IDF は一般語 (多くのドキュメントに出現する語) のフィルタとしての役割を果たす．

(4.2) 式において，Dは記事数であり，$\{d : d \ni w_i\}$ は固有名詞 w_i を含む記事数である．本研究では，w_i の朝日新聞，読売新聞での IDF 値をそれぞれ $IDF^F_{w_i}$，$IDF^E_{w_i}$ とする．固有名詞 w_i の IDF 値を以下の式を用いて計算する．

$$IDF_{w_i} = \alpha IDF^F_{w_i} + (1-\alpha) IDF^E_{w_i} \tag{4.3}$$

α は比例定数である．例えば，(4.3) 式において α=0.5 とし，表 4.2，表 4.3 中の固有名詞，**麻生**と**千葉**の IDF 値を計算すると，IDF(麻生)=0.5log2/1+0.5log2/1=log2，IDF(千葉)=0.5log2/1+0.5log2/2=0.5log2 となり，**麻生**という固有名詞の重みが**千葉**という固有名詞の重みより重くなる．(4.3) 式を用いて一致度 Mを以下のように再定義する．ただし固有名詞の IDF 値は (4.3) 式を用い，名詞の IDF 値は 1 [2]として計算する．

$$M_2 = \frac{\sum_i \sum_j IDF_{w^F_i = w^E_j} \times Match(w^F_i, w^E_j)}{max(\sum_i IDF^F_i \times C(w^F_i), \sum_j IDF^E_j \times C(w^E_j))} \tag{4.4}$$

本研究では (4.4) 式を用いて，すべての対応生成した記事の比較を行い，閾値を用いて対応記事中の本文の対応を得る．

4.4 類義語の抽出

本章では，前章で得られた文対応から統計翻訳手法を用いて類義語を抽出する方法を述べる．

4.4.1 統計翻訳手法による類義語抽出

本節では，本研究で用いた統計翻訳モデル [2] について述べる．統計翻訳では，ある起点言語 *F(Source Language)* からある目標言語 *E(Target Language)* への翻訳を，以下の式のように最尤翻訳文 $\hat{e}$ を求める問題に置き換える．

$$\hat{e} = \arg\max_e P(e)P(f|e) \tag{4.5}$$

実際に翻訳を行う際には，(4.5) 式において，$P(e)$ と $P(f|e)$ を求めることにより，翻訳文を生成する．ここで，$P(e)$ を言語モデル (*Language Model*)，$P(f|e)$ を翻訳モデル (*Translation Model*) とよぶ．本稿では得られた文対応から類義語を抽出するために必要な翻訳モデルに着目して述べる．

[2]実験から最も IDF 値の低い語が”東京”の 1.05 であったため，名詞の重みを 1 とした

翻訳モデルの代表的なものとして，Brown ら [4] による IBM Model がある．IBM Model には5つの定義（モデル）が存在するが，本研究では最も構造が単純な IBM Model1 を用いる．

IBM Model1 では，翻訳モデル $P(f|e)$ を計算するため，起点言語と目標言語の単語の対応位置関係をアラインメントとして定義する．起点言語を日本語，目的言語を英語とした図 4.1 において各単語を繋いでいる線がアラインメントを表現している．

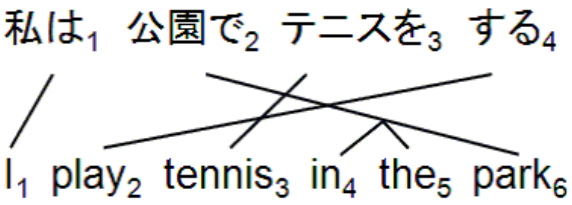

図 4.1: 日本語-英語のアラインメント例

アラインメント a を用いて $P(f|e)$ を以下のように定義する．

$$P(f|e) = \sum_{a} P(f,a|e) \tag{4.6}$$

$$P(f,a|e) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{4.7}$$

ϵ は比例定数であり，m は起点言語文の長さ，l は目標言語の長さ，f_j は起点言語文で j 番目に位置する単語，e_i は目標言語文で i 番目に位置する単語を表す．アラインメント a_j は起点言語文で j 番目に位置する単語が目標言語文で対応する単語の位置を表す．例えば，図 4.1 において，$a_4 = 2$ となる．

(4.7) 式において，求めるべきパラメタは $t(f_j|e_{a_j})$ であり，$t(f_j|e_{a_j})$ が単語 e_{a_j} が単語 f_j に翻訳される語彙確率を表す．パラメタ $t(f_j|e_{a_j})$ の推定は，対訳コーパスを用いて，ラグランジュの未定乗数法により求める．詳細は [4] を参照されたい．

4.4.2 類似度に対する重み付け

本研究では，朝日新聞と読売新聞を用いた IBM model1 をベースラインとし，領域依存型類語語の抽出を行う．対応生成した2つのニュース記事文の片方を起点言語，もう一方を目標言語として語彙確率を得る． ここで，語彙確率 $P(f_j|e_{a_j})$ を単語 e_i に対する単語 f_j の類似度と考える．例えば，(a1) と (d1) を用いて動詞の類似度を計算すると以下のようになる．

P(発表する | 立候補する) = 0.5
P(出馬する | 立候補する) = 0.5

IBM Model1 では語彙確率を対訳コーパスの起点言語・目標言語のアラインメント頻度から求めるため，出現しやすい単語の語彙確率が高くなりがちである．この問題を回避するため出現しやすい単語の重みとして IDF 値を用いる．前章で，単語 w_i の IDF 値を (4.3) 式のように定めたが，本章では，語彙確率 $t(f_j|e_{a_j})$ の重みとして用いるため，起点言語 f_j の IDF 値が重要となる．単語 f_j の IDF 値である $IDF^F_{f_j}$ を重みとして用いた類似度 $Sim_{e_i}(f_j)$ を以下のように定義する。

$$Sim_{e_i}(f_j) \quad = \quad t(f_j|e_{a_j}) \times IDF^F_{f_j} \tag{4.8}$$

また，本研究では共起性に関する重みも用いる．先ほどと同様に，IBM Model1 は語彙確率をアラインメント頻度から求めるため，共起しやすい単語の語彙確率が高くなりがちである．例えば，4.1 節で用いた t(発表する | 立候補する) $= 0.5$，t(出馬する | 立候補する) $= 0.5$ の場合を考える．(d1) 中の語の共起頻度を求めると，**立候補する**と**発表する**は共起している (共起頻度=1) が，(d1) 中に**出馬する**は存在しないため**立候補する**と**出馬する**は共起していない (共起頻度=0)．語彙確率の重みとして共起逆頻度を用いると，t(発表する | 立候補する) $= 0.5 \times 0 = 0$，t(出馬する | 立候補する) $= 0.5 \times 1 = 0.5$ となり (d1) 中の語**立候補する**の類義語を**出馬する**にできる．目標言語中の単語 e_i と単語 $e_{i'}$ の共起頻度 $k(e_{i'}|e_i)$ を条件付き確率によって以下のように求める．

$$k(e_{i'}|e_i) \quad = \quad \frac{C(e_{i'}, e_i)}{C(e_i)} \tag{4.9}$$

共起頻度を重みとして使用するため，対数逆頻度を用いて，類似度 $Sim_{e_i}(f_j)$ を以下のように定義する．

$$Sim'_{e_i}(f_j) \quad = \quad t(f_j|e_{a_j}) \times IDF^F_{f_j} \times (-logk(f_j|e_i)) \tag{4.10}$$

実験では，(??) 式をベースラインとし，(4.8) 式，(4.10) 式と用いる重みを変更して類似度を決定し，類義語の抽出を行う．

4.5 実験

本研究では，2 つの実験を行う．実験 1 では，ニュース記事から本文対応生成を行い，実験 2 では，得られた文対応を用いて類義語の抽出を行う．本研究では類義語の抽出を行う単語の品詞を動詞に限定する． 本研究の手法では記事の対応生成を行う際に，名詞または固有名詞の一致度を用いて対応生成を行うため，名詞と固有名詞に関しては類義語を得ることが困難であることに注目したい．

4.5.1 準備

本研究では，新聞コーパスとして朝日新聞 2007 年版と読売新聞 2007 年版を用いる．記事数として，朝日新聞には 153246 件，読売新聞には 343142 件の記事を用いる．月毎の内訳を表 4.4 に示す．また，テストデータとして手作業で対応生成した 350 行を用いる．

月	朝日新聞記事数 (件)	読売新聞記事数 (件)
1 月	11989	26795
2 月	12038	26567
3 月	13483	29647
4 月	12823	28437
5 月	13471	29331
6 月	13563	30092
7 月	12009	27912
8 月	11973	29294
9 月	12809	28814
10 月	13446	30200
11 月	13169	28569
12 月	12473	27484
合計	153246	343142

表 4.4: 朝日新聞，読売新聞の月別記事数

日本語形態素解析ツールとして MeCab[12] を用いる．IBM Model1 を学習する際には，本研究では起点言語を朝日新聞，目標言語を読売新聞とし，繰り返し回数を 5 回とする．類似度の重みに用いる IDF 値と共起頻度は，得られた文対応から動詞のみを取り出し計算する．

4.5.2 評価方法

本研究では，統計翻訳の評価関数である Word Error Rate (WER)[13] と，Position independent word Error Rate(PER) を用いる．

WER は語順を考慮した不一致率であり，参照訳との編集距離によって計算する．

$$WER = \frac{\sum_i (\text{挿入語数}\, i + \text{削除語数}\, i + \text{置換語数}\, i)}{\sum_i \text{参照訳}\, i \,\text{の語数}} \tag{4.11}$$

PER は語順を無視し、文を単語集合とした不一致率である．PER を用いることで，WER では翻訳が正しくても正解と語順が著しく異なる場合に結果が悪くなってしまうため，語順を考慮せず，語のレベルでその翻訳が正しいかを判断することができる．

$$PER = 1 - \frac{\sum_i (\text{翻訳文 } i \text{・参照訳 } i \text{ 間の一致語数})}{\sum_i \text{参照訳 } i \text{ の語数}} \tag{4.12}$$

ともに誤り率であるので，数値が低いほど良い結果となる．

4.5.3 実験結果

実験 1

ニュース記事の文対応生成に関する結果を示す．始めに、記事の見出し対応を一致度 M を用いて閾値を変化させた場合の記事数を表 4.5 に示す．また対応記事中の朝日新聞，読売新聞の本文数を表 4.6 に示す．

M_1	0.5	0.6	0.7	0.8	0.9	1.0
1 月	5538	2021	791	446	158	137
2 月	7097	3291	1029	519	185	162
3 月	8656	3444	1289	733	283	245
4 月	34344	16170	5859	3002	1209	922
5 月	7299	3028	1194	768	327	295
6 月	5878	2534	1027	603	256	221
7 月	17041	5143	1667	843	316	251
8 月	5977	2345	852	475	189	166
9 月	8957	3806	1367	723	287	240
10 月	9029	4835	1731	948	373	332
11 月	8537	63594	1099	631	205	185
12 月	4895	2196	899	560	223	207
合計	123248	52407	18804	10251	4011	3363

表 4.5: 実験結果:ニュース記事見出し月別対応記事数 (件)

どの閾値においても 4 月の対応件数が最も多くなっている．また，一記事あたりの平均本文数は，朝日新聞で 49.1 文，読売新聞で 39.7 文であった．

次に，一致度 M を用いて閾値を変化させた PER，WER を表 4.7 に示す．本来，記事同士が対応するかどうかはその見出しの要旨が同じどうかで判断すべきであるが，定量的な実験が困難なため，本研究では朝日新聞を翻訳文，読売新聞を参照訳として WER と PER を用いて評価を行う．

閾値	朝日新聞本文数 (文)	読売新聞本文数 (文)
0.5	7160307	5963342
0.6	3145875	2507245
0.7	1037026	833717
0.8	460683	375253
0.9	161648	132916
1.0	120703	92837

表 4.6: 実験結果:対応記事中の本文数

閾値	PER	WER
0.5	0.57	0.94
0.6	0.50	0.87
0.7	0.43	0.80
0.8	0.40	0.74
0.9	0.41	0.70
1.0	0.42	0.67

表 4.7: 実験結果:ニュース記事見出しの PER，WER

最も PER の値が良いのは $M_1 = 0.8$ としたときであり，PER の値は 0.40 である． 最も WER の値が良いのは $M_1 = 1.0$ としたときであり，WER の値は 0.74 である．

次に，$M_1 = 0.8$ を用いて見出しの対応生成を行った記事から，本文対応生成を一致度 M_2 を用いて，閾値を変化させた場合の結果を表 4.8 に示す．

閾値	PER	WER
0.5	0.34	0.86
0.7	0.35	0.84
0.9	0.36	0.82

表 4.8: 実験結果:ニュース記事本文の PER，WER

最も PER の値が良いのは $M_2 = 0.5$ としたときであり，PER の値は 0.34 である． 最も WER の値が良いのは $M_2 = 0.9$ としたときであり，WER の値は 0.82 である．

実験 2

類似度 $P(e_i|f_j)$，$Sim_{e_i}(f_j)$ and $Sim'_{e_i}(f_j)$ を用いて類義語の抽出を行った結果を示す． 記事見出しの一致度を $M_1 = 0.8$，記事本文の一致度を $M_2 = 0.9$ として得られた

記事本文対応を用いる． 例として**立つ**と**打つ**に関する類義語の抽出を行った結果の類似度が高い順に並べた上位 10 件を表 4.9，表 4.10 に示す．

順位	Model1	Sim	+IDF	Sim	+共起頻度	Sim
1	なる	0.086	立つ	0.086	立候補する	0.088
2	立候補する	0.081	立候補する	0.068	立つ	0.086
3	立つ	0.049	推薦する	0.051	巡る	0.080
4	告示する	0.032	締め切る	0.040	出馬する	0.076
5	推薦する	0.031	巡る	0.039	届け出る	0.069
6	巡る	0.026	なる	0.035	進める	0.068
7	決まる	0.024	出馬する	0.034	出席する	0.056
8	出馬する	0.022	決まる	0.034	スタートする	0.056
9	決める	0.022	進める	0.031	推薦する	0.051
10	届け出る	0.021	スタートする	0.030	向ける	0.050

表 4.9: 実験結果:類義語の抽出 (立つ)

順位	Model1	Sim	+IDF	Sim	+共起頻度	Sim
1	打つ	0.354	打つ	0.721	打つ	0.721
2	死亡する	0.127	来る	0.243	来る	0.288
3	来る	0.108	死亡する	0.234	運ぶ	0.221
4	調べる	0.104	調べる	0.181	調べる	0.210
5	運ぶ	0.063	運ぶ	0.153	死亡する	0.184
6	はねる	0.048	はねる	0.102	はねる	0.097
7	する	0.029	衝突する	0.062	衝突する	0.078
8	衝突する	0.024	乗る	0.038	乗る	0.064
9	よる	0.019	右折する	0.031	右折する	0.063
10	乗る	0.016	渡る	0.030	渡る	0.050

表 4.10: 実験結果:類義語の抽出 (打つ)

表 4.9，表 4.10 から提案手法は**立つ**という語に対して，**立候補する**，**出馬する**という類義語を，**打つ**という語に対して，**はねる**，**衝突する**という類義語を抽出可能である．

最後に得られた類義語の評価を行う． 類義語の評価は，単語の意味が類似性で決定するため定量的な評価は行えない． 手作業で対応生成したテストコーパスを用い，抽出した類義語を用いて朝日新聞コーパスを書き換える前後での PER，WER の値の比較することにより評価を行う． 類義語を最も類似度が高い語で置き換えた場合の PER，WER の値を表 4.11 に示す．

	類義語による書き換え前	類義語による書き換え後
PER	0.69	0.63
WER	0.92	0.91

表 4.11: 実験結果:類義語を用いる前後での PER，WER の比較

PER の値で 6%の改善が見られる．

4.5.4 考察

実験 1 に関する考察を行う． 記事見出しの対応生成では，閾値を $M_1 = 0.8$ としたときに最も PER の値が良く，0.40 である．このとき、WER の値は 0.74 である．見出しは文の形をなしていないものが多いため，PER の値が良いことが望ましい．したがって閾値 0.8 では，6 割の同じ単語を含むが，文としては語順が異なる見出しが対応可能である．なおかつ多くの同じ固有名詞を含むので，同じ趣旨の見出しが対応可能であると考える．例えば，以下のような見出しの対応が得られた．

- 朝日新聞:“「信頼し合って暮らす社会に」天皇陛下が感想 ”
- 読売新聞:“ 天皇陛下が新年の感想「皆が信頼して暮らせる社会を」”

上記の例のように，同じ単語を多く含むが，語順が異なる見出しの対応生成できる．また，対応記事数としては閾値によらず 4 月が最も多かった．これは 4 月に統一地方選挙が行われたため，通常 1 対 1 で記事が対応するはずであるが，複数の地域で同様な見出しが使われ，多対多で対応してしまったためであると考えられる．実際，以下に示すような見出しを対応可能と判定してしまっている．

- 町村議選の候補者　統一地方選　／山梨県
- 町村議選の候補者　統一地方選　／青森県
- 町村議選の候補者　統一地方選　／千葉県
- 町村議選の候補者　無投票当選者　統一地方選　／長野県

これらの見出しは，新聞記事の地域欄によるものであると考えられる．

得られた記事対応からの本文の対応生成では，閾値を $M_2 = 0.9$ としたときに最も WER の値が良く，0.82 である．見出しの対応とは異なり，本文の対応であるため，文として類似することが望ましいので，WER の値が良いことが望ましい．

例えば，以下のような本文の対応が得られている．

- 朝日新聞:“ 第８５回全国高校サッカー選手権は２日、２回戦で県代表・秋田商が神村学園（鹿児島）と対戦し、ＰＫ戦で惜敗した 。 ”

- 読売新聞:“ 全国高校サッカー選手権大会（読売新聞社など後援）は２日、２回戦１６試合が行われ、県代表の秋田商は、横浜市の三ツ沢球技場で神村学園（鹿児島）と対戦し、０－０の同点で迎えたＰＫ戦で惜しくも敗れ、初戦突破はならなかった 。 ”

上記の例のように，語順が類似するが，動詞が異なる文の対応生成が可能である．

文対応の PER の値は，見出し対応の PER の値より良いが，これは PER が単語集合としての類似度を示す尺度であり，見出しに比べ本文は長いため，より多くの単語を含んでいるからであると考える．それに伴い，文対応の WER の値は，見出し対応の WER の値より悪化している．

実験２の考察を行う．表 4.9，表 4.10 から分かるように，**なる**や**する**といった出現しやすい語の類似度を下げることが可能である．例えば，**なる**という単語は，得られた文対応 7629 文中 2965 文に含まれる．したがって，その IDF 値は 0.41 となるので，類似度を下げることが可能である．

また，**打つ**という語に比べ，**立つ**は上位に類義語が抽出可能である．これは 2007 年に，参議院議員選挙や統一地方選挙があり，選挙に関連する記事が多く出現していたからであると考える．実際に，**立つ**は 7629 文中 140 文，**打つ**は 7629 文中 70 文と，**立つ**を含む文の方が多く見られた．

最後に，類義語を適用する前後での PER，WER に関する実験についての考察を行う．類義語を用いることで，PER が 6%改善している．例えば，

- 朝日新聞:“ 市は滞納額の上位１００人のうち、支払いの意思表示がない７４人に通知書を送付した。”

- 読売新聞:“ 市保育運営課によると、９月と１０月、滞納額の多い上位１００人のうち、再三の 督促にもかかわらず分割納付に応じず、納付する意思を示さない７４人に財産 差し押さえの事前通知書を郵送した 。 ”

という文対に対し，朝日新聞の動詞に類義語を適用すると，

- 朝日新聞:“ 市は滞納額の上位１００人のうち、支払いの意思表示がない７４人に通知書を郵送した。”

となる．この例文では，単語対応が取れるようになるため PER が上昇している．しかし，例文間で単語数が違うため，語順が異なるので WER の改善は難しい．全体として PER が 6%改善していることから，類義語を用いることにより朝日新聞の動詞を読売新聞の動詞に書き換えることが可能である．したがって，領域依存類義語が抽出可能であると考える．

4.6 結論

本研究では，類義語を自動的に抽出する方法として，統計翻訳モデルを用いた方法を提案した． その際に，複数のニュースコーパスの対応生成を行う方法を述べ，文対応の生成も自動的に行った． 得られた文対応から統計翻訳モデルを用いて動詞の類義語自動抽出を行った． 実験により得られた類義語を対訳文に用いることで6%の精度改善が見られたことから提案手法が領域に依存した類義語を抽出可能であることを確認した．

第5章　結論

本研究では，統計機械翻訳における問題を解決するための領域に依存した情報の抽出を行う手法の提案を行った．

まずはじめに，形態素解析中の固有名詞の認識を考えた．学習データからHMMモデルを構成し，固有名詞認識のためのルールの自動抽出を行った．その後HMMによって品詞を推定したテストデータに対してルールを適用した．また，抽出したルール集合を改良するため，品詞フィルタリングの手法を提案し，実験から手法が有用であることを確認した．

次に，上記の手法による形態素解析を用いて統計翻訳におけるアラインメント問題を考えた．英語と日本語のフレーズアラインメントを生成する方法として，形態素解析を用いたフレーズ生成を行い，得られたフレーズに対し，隠れマルコフモデルを用いるアラインメント手法を提案した．実験結果から提案手法がフレーズアラインメントを生成する方法として期待できることを確認した．

最後に，類義語を自動的に抽出する方法を考えた．ここでは領域に依存した類義語に注目し，統計翻訳モデルを用いた方法を提案した．その際に，複数のニュースコーパスの対応生成を行う方法を述べ，文対応の生成も自動的に行った．得られた文対応から統計翻訳モデルを用いて動詞の類義語自動抽出を行った．実験により得られた類義語を対訳文に用いることで6%の精度改善が見られたことから提案手法が領域に依存した類義語を抽出可能であることを確認した．

本研究で提案した手法で得られる領域依存情報は，学習データに用いた領域に依存した解析結果を表現することができる．統計翻訳で示したように，正しい固有名詞情報は翻訳精度に大きな影響を与えていた．また，文書分類においては，あらかじめ提案手法を用いて類義語の抽出を行っておき，得られた類義語を文書分類手法に適用することで，一般的な類義語を用いるよりも高精度に分類を行えることが期待できる．このように，提案手法は幅広い分野への応用が期待できる．

今後の課題として，類義語抽出における領域の細分化が挙げられる．本研究では領域の単位を**ニュース**としたが，例えば，**スポーツ**，**政治**などと細分化することでより高精度な領域依存類義語抽出を行えるだろう．

また，得られた類義語を文書分類や情報検索などの他分野の研究に適用し，一般的な類義語を用いた場合との結果の比較を行い，領域依存型情報が有用であるか確認を行う必要がある．

謝辞

本研究を遂行するにあたり，日頃より適切な御指導，御鞭撻をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます．

並びに，産業能率大学情報マネジメント学部 塩谷勇教授にも多くの有益なご助言をいただきました．深く感謝いたします．

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適で能率的な研究室環境を整えていただきました．御礼申し上げます．

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます．

最後に，今までの学生生活を支えてくださった私の両親，兄弟に深く感謝したいと思います．

参考文献

[1] Abney, S.: Part of Speech Tagging and Partial Parsing, In *Corpus-Based Methods in Language and Speech*, Kluwer Academic Publishers, 1996

[2] Adam Lopez, Statistical Machine Translation, ACM Computing Surveys, Vol. 40, No. 3, Article 8, Publication date: August 2008.

[3] Bernard Merialdo: Tagging English Text with a Probabilistic Model. *Association for Computational Linguistics* pp. 155-171, 1994

[4] Brown, P. F., Pierta, S. A. D., Pierta, V. J. D., and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist. 19, 2, 263.311

[5] Brown Corpus: www.essex.ac.uk/linguistics/ clmt/w3c/corpus_ling/content/ corpora/list/private/brown/brown.html

[6] George A. Miller. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[7] Heng Ji. and Ralph Grishman: Analysis and Repair of Name Tagger Errors. *Proceedings of the COLING/ACL* pp. 420-427, 2006

[8] Ian Fette: Combining n-gram based statistics with traditional methods for named entity recognition. Carnegie Mellon University, 2007

[9] Kyoko Kanzaki, Francis Bond, Noriko Tomuro and Hitoshi Isahara. Extraction of Attribute Concepts from Japanese Adjectives. 2008.

[10] Manning, C.D. and Schutze, H.: Foundations of Statistical Natural Language Processing, MIT Press, 1999

[11] Masao Utiyama and Hitoshi Isahara. Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72--79.

[12] MeCab: http://Mecab.sourceforge.net/

[13] OCH, F. J., TILLMAN, C., AND NEY, H. 1999. Improved alignment models for statistical machine translation. In Proceedings of EMNLP-VLC. 20.28.

[14] Ryohei AGEISHI, Takao MIURA. Named Entity Recognition Based On A Hidden Markov Model in Part-Of-Speech Tagging, ICADIWT, 2008.

[15] 揚石 亮平，三浦 孝夫．形態素解析とHMMを用いたフレーズアラインメント，データ工学と情報マネジメントに関するフォーラム (DEIM)，2009．

[16] 揚石 亮平，三浦 孝夫．統計翻訳手法を用いた類義語の自動抽出，データ工学と情報マネジメントに関するフォーラム (DEIM)，2010．

[17] 浅原 正幸:系列ラベリング問題に関するメモ，奈良先端科学技術大学院大学，2006

[18] 英辞郎:100 万語収録付録 CD-ROM，株式会社アルク

[19] 北 研二:確率的言語モデル，東京大学出版会，1999

[20] 北 研二，津田 和彦，獅々堀 正幹，情報検索アルゴリズム，2002，共立出版

[21] 中渡瀬 秀一，複合語からの類義語抽出法，2002，情報処理学会研究報告 pp.39-46