事例に基づく英語動詞選択ルールの修正型学習手法

金田 重郎，秋葉 泰弘，石井 恵
NTTコミュニケーション科学研究所
E-mail:{kaneda,akiba,megumi}@nttkb.ntt.jp

1 はじめに

本稿では，英語動詞選択ルールの獲得問題を取り上げる．ここで，英語動詞選択ルールとは，日本語動詞を英語動詞に翻訳する際に，英語動詞を決めるルールであり，機械翻訳システムでは，極めて多数の英語動詞選択ルールを必要とする．このため，その自動作成/作成支援技術が求められている．

Almuallimらは，機械学習アルゴリズムを利用し，日英翻訳事例から，英語動詞選択ルールを自動生成する手法を提案した[2]．この従来手法では，ある日本語動詞に対する英語動詞選択ルールを獲得するためには，一定個数の翻訳事例を必要とする．

しかし，既存の文書から十分な個数の翻訳事例を収集できる動詞の種類は限られる．従って，事例があまり手に入らない日本語動詞についても，英語動詞選択ルールを自動生成するには，現実の事例の少なさを補う何らかの情報を準備し，その情報と現実事例の融合により，精度の高いルールを得る必要がある．

そこで，本稿では，人手作成の既存ルールを利用し，これと事例を組み合わせてルールを獲得する，知識修正型の学習手法を提案する．本手法の特徴は，(1)既存ルールから事例(以下"**仮事例**"と呼ぶ)を生成し，この仮事例と，実際の事例(以下"**実事例**"と呼ぶ)とを併せて訓練事例として利用する，(2)仮事例と実事例に，それぞれどの程度の重みを置いて学習をすべきかをクロスバリデーション法により判定する，の2点である．

既存文書を参考にして作成した翻訳事例と人手作成のルールを用いて本手法を実験的に評価したところ，本提案の手法で生成されるルールは，翻訳事例のみから生成されたルールや人手作成ルールより高い正解率を示した．

IF		THEN	
J-Verb	= "焼く"		
N_1 (主格)	≡"人"	E-Verb	= "bake"
N_2 (目的格)	≡"パン" or "菓子"		

図1: "焼く"の英語動詞選択ルールの例

2 英語動詞選択ルールとその獲得問題

NTTで開発中の日英機械翻訳システムALT-J/Eでは，日本文パターンと英文パターンの対応関係を規定する翻訳ルールを用いて日英翻訳を行っている．英語動詞選択ルールは，この翻訳ルールの主要部をなし，一つの日本語動詞に対して，複数の英語動詞を対応づける(図1参照)．

図1に示す様に，英語動詞選択ルールは，ルール条件部に日本文パターンを，ルール実行部に英語動詞を持つ．ここで，日本文パターンは，一つの日本語動詞，格要素(『主格』，『目的格』等)の主名詞が持つべき意味属性条件から構成される．N1，N2等はこの格の種別を示す．[魚]，[魚介類]等は，意味属性であり，ALTでは，約3000個に及ぶ意味属性が階層的なシソーラスを形成している．

英語動詞選択ルール構築において最も困難なタクスは，ルール条件部における格要素のとるべき意味属性の決定である．他のルールと相互矛盾なく，ルール条件部の意味属性を決めることは，意味属性体系に熟知した専門家であっても，多大の労力を要する．

3 従来手法とその問題点

英語動詞選択ルールの自動生成技術としては，Almuallimらによる機械学習アルゴリズムを利用した手法がある[2]．この従来手法は，意味属性シソーラスと次のよう訓練事例を前提として，英語動詞選択ルールの学習を実現している．

まず，”ターゲット日本語動詞を含む日本語の単文と英語動詞の対”を用意する．例えば，ターゲットの日本語動詞が”焼く”なら，

(”コックがアップルパイを焼く”　”bake”)

のような対を用意する．次に，各対の日本語部分をパーザに通し，動詞，主名詞に分解した結果を事例とする．

上記従来手法では，ターゲット日本語動詞について，十分な数(動詞にも依存するが最低100個程度)の事例が準備できれば，高い正解率をもつターゲット英語動詞選択ルールが生成可能であることが示されている．

そこで，十分な数の事例が準備できるか否かを調査するために，既存の文書を参考として，日英翻訳コーパス(単文)を約5万事例収集した．以下，得られたコーポラの主要分析結果をに示す．

分析結果1 十分に事例を用意出来る日本語動詞は限定される．100個以上の事例をもつ動詞は，全体の約1%程度にすぎない．なお，文中の全日本語動詞は，約5000個である．

分析結果2 多数の事例があっても，同一の文が頻出し，実質的に利用出来る事例数はさらに減少する．

分析結果3 出現頻度によりコーパス中の動詞に順位をつけ，上位から95%の動詞まで採ると，最も出現頻度の小さな動詞のコーパス中の出現回数は2回であった．各日本語動詞に対して，一英語動詞動詞当たり25個の翻訳事例が必要で，対訳英語動詞が4種類あるとすれば，一つの動詞に少なくとも100事例は欲しい．従って，単純計算すると，95%の動詞をカバーするには，少なくとも250万以上の翻訳事例を必要となる．

これらの現象を見るかぎり，事例さえ集めて統計的に処理すれば何でもできると考えるのは，楽観的にすぎる．十分に翻訳事例を得られない日本語動詞についても，英語動詞選択ルールを自動生成するには，実事例の少なさを補う何らかの情報を準備し，その情報と実事例の融合により，精度の高いルールを得る必要がある．

そこで，著者らは，人間の持つ英語動詞選択のための知識を，人手作成のルールとして取りだし，これと収集可能な事例を融合するアプローチを採用した．次節では，人手作成した英語動詞選択ルールと，翻訳事例とを総合する学習手法(以下，**既存ルール修正型学習手法**，または，単に**修正型学習法**と呼ぶ)を提案する．

4　既存ルール修正型学習手法

最初に，タスクを整理しておく．

【学習タスク】

(STEP I) 専門家は自らルールを作成する(作成されたルールは性能的に十分なものではない)．

(STEP II) 事例を収集する(収集された事例から学習手法によりルールを作成してもそのルールは，性能的に十分なものではない．また，人手作成のルールに合致するものもあれば，合致しないものもある)．

(STEP III) 既存知識修正型の機械学習アルゴリズムにより，上記の事例とルールから，最終的なルールを自動作成する．□

上記のタスクは，Theory Revisionとして，すでに多くの研究者により，研究されて来た．しかし，自然言語処理に適用できる，高速なアルゴリズムは，知られていない．

4.1　修正型学習手法

提案手法を以下に示す．

【既存知識修正型アルゴリズム】

(STEP1) 既存のルールを元にして事例を生成する．生成された事例を，本稿では，”**仮事例**”と呼ぶ．仮事例の具体的な生成方法は後述する．

(STEP2) 自然界，即ち，既存の文書等から収集された事例を”**実事例**”と呼ぶ時，この実事例と仮事

例とを併せて，訓練事例とする．

(STEP3) 意味シソーラスと上記の訓練事例とを既存の機械学習アルゴリズム(以下，これを**内部学習アルゴリズム**と呼ぶ)に入力して，最終的なルールを得る．□

なお，本稿では，事例は，プロパティ値(属性値)と，そのプロパティ値により決まるクラスから構成されるとする．仮事例は，以下の手法により生成される．

【仮事例生成アルゴリズム】

(Step i) 既存ルールを，"**単位ルール**"に分解する．単位ルールとは，プロパティの条件として，OR条件を持たないルールで，次式で表わせる．

IF $(N_1 = V_1)$ & $(N_2 = V_2)$ & ...

THEN Class =CV[1]

ここで，N_1，N_2 等はプロパティ名称，V_1，V_2 等はプロパティ値である．CV はクラス名称である．

(Step ii) 上記の単位ルールから，以下の形式の事例をランダムに生成する．

$$(v_1, v_2, ... : \mathrm{CV})$$

$v_i (i = 1, 2, ...)$ は意味シソーラス上の葉(最下位のノード)であってかつ，単位ルール中の各プロパティのプロパティ値 V_i の意味シソーラス上の下位ノードある．

(Step iii) 上記の Step2 を，予め定めた個数の事例が生成されるまで繰り返す．□

4.2 事例重みのチューニング

既存知識，即ち，仮事例が信用でき，これに対して実事例が信用できない場合には，仮事例に大きな重みをおいて，学習アルゴリズムを作動させるべきである．これに対して，事前確率である仮事例が信用できず，実事例の方が確かな場合には，実事例に重みをおくことで，良い精度が得られると考えられる．この重みは，予め与えられてはいないので，学習アルゴリズムが定める必要がある．

[1]否定が条件に現われる時には，(N1= not V1) と表現する．

そこで，本稿では，クロスバリデーション法により，重みを定めることとした．仮事例と実事例の重みを変えながら，クロスバリデーションを繰り返して，もっとも正解率の高い重みを得る．なお，クロスバリデーションにおけるテスト事例は，仮事例と実事例からの両方から選んでいる．

5 評価

以下に示す設定で，提案手法を実験的に評価した．対象の日本語動詞は，"入る"，"見える"，"見る"，及び"取る"で，人手作成された各日本語動詞に対する英語動詞翻訳ルールは，重複や抜けがあるため，完全なものではない．なお，各日本語動詞に対する英語動詞翻訳ルールの英語動詞数は，"入る"，"見える"，"見る"，"取る"の順に，4，1，3，8個である．実事例は，既存の文書を参考にして作成した．準備した実事例の数は，"入る"，"見える"，"見る"，取る"の順に，95，33，358，130個である．また，各日本語動詞に対する実事例中の英語動詞数は，27，4，40，46個である．各動詞に対する実事例の中には，準備した対応する英語動詞翻訳ルールに合致するものもあれば，合致しないものもある．

仮事例は，準備した実事例と同数を仮事例生成アルゴリズムで生成した．事例を表現する格は，実事例中にもっとも多く出現した格を必須格とみなして採用した．意味シソーラスは，NTTで研究開発中の機械翻訳システム ALT-J/E に採用されているものである．内部学習アルゴリズムとしては，C4.5[6] を利用した．但し，事例の重み自由に設定できる様に改造と加えている．クロスバリデーションは，10 fold cross validaiton である．

図2に，実験結果を示す．実験では，(1) 従来手法[2] に実事例を入力して生成されたルール，(2) 提案手法で生成されたルール，及び (3) 従来手法に仮事例を入力して生成されたルールの正解率を比較した．

○提案手法の正解率は，実事例のみを従来手法[2] に投入した場合と比べ，大幅に改善されている．また，仮事例を，従来手法に投入した場合に比べても遜色ない．これは，提案手法により生成された

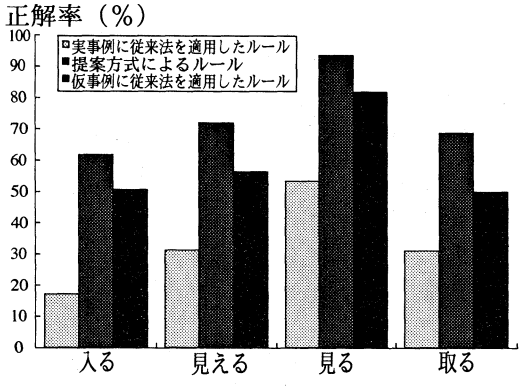

図 2: 実験結果

ルールにより，人手作成の既存ルールでは説明出来なかった実事例が説明される様になったことを示す．

○仮事例のみから生成されたルールの正解率は実事例のみで生成されたルールの正解率に比べて高い．一英語動詞動詞当たりの仮事例数が一英語動詞動詞当たりの実事例数より多いためと考えられる．特に，実事例のみでルールを生成した場合，訓練事例含まれていない英語動詞が，テスト事例が数多く含まれる可能性があり，そのため低い正解率に留まっている．

○提案手法による日本語動詞"見る"に対するルールだけが，90% を越える高い正解率を示し，実事例に従来法を適応したルールと仮事例に従来法を適応したルールも，他の動詞に対するそれより高い正解率を示している．これは，"見る"だけが，実事例に含まれる英語動詞と仮事例に含まれる英語動詞に含まれる英語動詞に共通な項が多いためである．

これらから，以下の事が結論出来る．提案手法は，従来手法より高い正解率を示す英語動詞選択ルールを生成する．また，絶対的に正解率を上げるためには，実事例に含まれる英語動詞全てについて，その英語動詞を結論部とする人手作成の英語動詞翻訳ルールを準備すればよいと推測される．今後は，この推測を検証する予定である．

6 おわりに

本稿では，人手作成の荒いルールと事例とを融合して，より精度の高いルールを獲得する修正型学習手法を提案した．具体的には，既存ルールから仮事例を生成し，この仮事例と実事例とを併せて内部学習アルゴリズムに対する訓練事例とする．この場合，仮事例に対する実事例の重みの決定が課題となるが，クロスバリデーションによる最適重み決定法を用いた．

内部学習アルゴリズムとして，高速な機械学習アルゴリズム C4.5 を利用して，機械翻訳システム ALT-J/E の動詞選択ルールを学習する実験を行なった．その結果，既存ルール / 実事例の何れからよりも高い性能を持つルールを獲得できた．本手法の最大の特長は，事例の表現が属性型を用いるアルゴリズムであれば，任意の学習アルゴリズムを利用出来る点にある．従って，本手法は，他の学習アルゴリズムを，修正型に拡張する手段としても，効果的である．

参考文献

[1] 秋葉，et al.：”帰納学習による日本語動詞翻訳ルールの自動獲得”，情報処理学会秋期全国大会，2k–06，(1994).

[2] Almuallim, Akiba, Yamzaki, Kaneda : "Two Methods for Learning ALT-J/E Translation Rules from Examples and A Semantic Hierarchy", Coling 94 , pp.57-63, (1994).

[3] Ikehara,S.,Shirai, S., Yokoo, A. and Nakaiwa, H.: "Toward an MT System without Pre-Editing-Effects of New Methods in ALT-J/E", Proc. of MT Summit-3, (1990).

[4] 池原，宮崎，横尾：“日本語機械翻訳のための意味解析辞書”，電子情報通信学会，研究会報告，NLC 91-19, (1991).

[5] 金田，Almuallim，秋葉，山崎：”意味属性体系を用いた事例に基づく翻訳ルールの学習”，「自然言語と実働」シンポジウム論文集 PP.144-150，電子情報通信学会，日本ソフトウエア科学会(1993)

[6] Quinlan : " C4.5 Programs for Machine Learning" Morgan Kaufman, San Mateo, California, (1992).