帰納学習による日本語動詞翻訳ルールの自動獲得

2K－6

秋葉 泰弘　石井 恵　金田 重郎
NTTコミュニケーション科学研究所

1 はじめに

帰納学習による日本語動詞翻訳事例からの日本語動詞翻訳ルールの獲得問題を取り上げる。日本語動詞を英語動詞に翻訳する際には、複数の訳が可能なため、日本語動詞翻訳事例は、従来の帰納学習アルゴリズムの訓練事例としては、そのままでは、適さない。そこで、日本語動詞翻訳事例から帰納学習に適した訓練事例を自動生成し、その訓練事例から帰納的学習アルゴリズムを利用して日本語動詞翻訳ルールを自動獲得する手法を開発した。本手法を、辞書から抽出した日本語動詞翻訳事例に適応したところ、高い正解率を示すルールを獲得できた。

2 タスク

実用的な、知識に基づく機械翻訳システムを構築するには、非常に多くの知識を準備する必要がある。例えば、NTTで研究・開発中の機械翻訳システムALT-J/E [Ikehara 90]の場合は、約13,000個のルールを持つ。一般的に、巨大なルールベースの作成・メンテナンスは、非常に難しい。また、知識に基づく機械翻訳システムの場合は、翻訳する文章の対象分野によって、知識ベースを改良する必要があるが、これも、同様に難しい。そのため、翻訳知識を自動獲得する方法論が、求められている。

知識を自動獲得する試みとしては、機械学習の帰納学習がある。帰納学習のアルゴリズムは、訓練事例を入力とし、それぞれの訓練事例を説明するルールを出力する。訓練事例は、訓練事例の属性を成分とするベクトルとその訓練事例の所属すべきクラスの対で表現される。出力されるルールは、どのような属性値(事例が持つ属性の値)を事例が持てば、どのクラスに所属すべきかを示唆する。

本稿のタスクは、日本語動詞翻訳事例(以下、翻訳事例と呼ぶ。)を訓練事例とし、帰納学習アルゴリズムにより、日本語動詞翻訳ルール(日本語動詞をどの英語動詞に翻訳すべきかを示すルール)を獲得することである。訓練事例の属性は、日本文を構成する格とし、属性値は、その格の名詞が持つ意味とし、クラスはその日本文を翻訳する際に選択した英語動詞とした。例えば、"女王が金を使う"なる日本文を、"使う"をspendに翻訳する事例として、次のように表現する。

$$\left\langle \left[\begin{array}{l} \text{主格} \equiv \{\text{貴族, 娘, 女性}\} \\ \text{目的格} \equiv \{\text{財産, 金属, 曜日, メダル}\} \end{array} \right], spend \right\rangle \quad (1)$$

ところで、帰納学習のアルゴリズムが求める訓練事例の集合に関する条件の中に、"クラスオーバーラップがない"ことが上げられる。クラスオーバーラップとは、訓練事例を表現する属性値が全て一致するが、クラスが異なる事例が含まれる現象である。一方、日本語動詞を英語動詞に翻訳する際には、複数の訳が可能なため、訓練事例の集合にクラスオバーラップが起こり得る。

以下、このクラスオバーラップのある訓練事例からルールを生成する方法を報告する。

3 アプローチ

本稿で提案するアプローチは、Almuallimらが提案した、クラスオーバーラップがない翻訳事例から日本語動詞翻訳ルールの自動獲得手法[Almuallim 94]をクラスオーバーラップがある翻訳事例の場合へ拡張したもので、次の処理を順に行なう。

1) クラスオーバーラップの問題のある訓練事例からクラスオーバーラップのない訓練事例を生成。

2) 属性値に多価を許容する事例を各属性の属性値が一価の事例へ変換。

3) 生成した事例を従来の事例に基づく学習アルゴリズムに入力。

以下、1)の事例の変換方法の詳細を述べる。2)の詳細は、文献[Almuallim 94]を参照のこと。3)は、帰納学習アルゴリズムであれば、何でもよい。

まず、次のような形式の属性ベクトルで表現された、ターゲットとしている日本語動詞の翻訳事例が与えられたとする。

$$\langle [N_1 \equiv S_1, \cdots, N_i \equiv S_i, \cdots, N_n \equiv S_n], E\text{-}Verb \rangle$$

---

†Acquisition of Translation Rules Using Machine Learning Technique.

Yasuhiro Akiba , Megumi Ishii, Shigeo Kaneda
NTT Communication Science Laboratories,
1-2356, Take, Yokosuka-shi, Kanagawa-ken, 238-03, JAPAN

ここで、この属性ベクトルは、日本語の単位文と対応する正しい英語動詞の対を表現している。N_i は、主語や目的語の様な日本文の構成要素を表現しており、"$N_i \equiv S$" は、日本文の N_i 成分の名詞が、$s \in S$ なる意味 s のインスタンスであることを意味する。なお、S は、空集合でない。これは、主語や目的語等の日本文の構成要素は、省略されず、かつ収集された翻訳事例は、同じ文型であることを意味する。

手続 1-1: 翻訳家が、英語動詞選択に本質的に重要であると判断した属性からなる属性空間上に、翻訳事例を射影。

手続 1-2: 射影空間上で、ハミング距離が 0 である事例同士をクラスタリング。

手続 1-3: 各クラスター毎に、そのクラスターに属する翻訳事例のクラスのうち、いずれのクラスが最瀕のクラスであるかを調べ、その最瀕のクラスでクラスター内の全事例のクラスの付け変え。

4 評価と考察

本稿の提案手法が、日本語動詞翻訳ルールを作成支援する上で、どの程度実用的であるかを評価するために、以下の実験を行なった。

まず、翻訳事例を生成するために、辞典([Keene 91]、[文化庁 90]) から日英対訳コーポラを抽出し、ALT-J/E のパーザー等を利用して、日英対訳コーポラを構文解析し、その結果を利用して、翻訳事例を自動生成した。抽出した日英対訳コーポラは、約 48,000 対で、その中に含まれている日本語動詞は、5,346 動詞であった。100 以上の翻訳事例をもつ 61 の日本語動詞を自動獲得のターゲット日本語動詞とし、その動詞の最瀕の文型をもつ翻訳事例を実験に利用した。

上記の様にして得たオーバーラップのある各々の翻訳事例に対して、本稿の提案手法と従来手法 [Almuallim 94] を適応した。各実験は、10 fold cross validation により、正解率を計っている。cross validation のための訓練事例は、翻訳事例集合を 10 個に分けその内の 9 個を和集合を取ったもので、テスト事例は、残りの部分集合である。正解率は、上記のようにして生成される訓練事例とテスト事例の 10 組の平均である。また、手続き 1-3 で使用する帰納学習アルゴリズムは、Quinlan の C4.5[Quinlan 93] をオプションなしで利用した。

実験結果を、表 1 に示す。いずれの場合も、本稿の提案手法の方が、従来手法より高い正解率示した。しかし、約 48,000 対程度の日英対訳コーパスで、各文型について十分な事例が集められたのは、全日本語動詞の 1% に過ぎない。全てこの方法によりルールを獲得するには、10 億対程度の日英対訳コーポラが必要である。従って、日英対訳コーパスの収集支援、より少ない事例で学習が可能な帰納学習アルゴリズム、人手で作成されたルールを精錬する学習アルゴリズム、いずれかの技術が必要がある。

表 1: 評価結果

日本語動詞	事例数 (個)	正解率 (%)	
		提案手法	従来手法
話す	76	89.5	84.3
作る	167	73.6	52.2
飲む	159	98.8	80.5
引く	94	82.1	45.9
入れる	77	71.6	74.3
受ける	91	67.2	34.3
平均		80.5	61.9

5 おわりに

本研究では、日本語動詞翻訳事例からの日本語動詞翻訳ルールを獲得する手法として、日本語動詞翻訳事例から帰納学習に適した訓練事例を生成し、その訓練事例から帰納的学習アルゴリズムを利用して日本語動詞翻訳ルールを自動獲得する手法を示した。辞書から抽出した日英対訳コーポラから生成した日英翻訳事例から本手法により日本語動詞翻訳ルールを生成したところ、高い正解率を示すルールを獲得できた。しかし、一部の日本語動詞に対して、日英対訳コーパスの絶対量が不足しているため、今後は、この日英対訳コーパスの不足を解決する技術が必要である。

参考文献

[Almuallim 94] H. Almuallim, Y. Akiba, T. Yamazaki, S. Kaneda, "Induction of Japanese-English Translation Rules from Ambiguous Examples and a Large Semantic Hierarchy.", 人工知能学会誌, 9(5), 1994.

[Ikehara 90] 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.

[Quinlan 93] Quinlan, J. R. *C4.5: programs for machine learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[Keene 91] Keene, D. 等, 会話作文 英語表現辞典 *(新訂版)*, 朝日出版社, 1991.

[文化庁 90] 文化庁, 外国人のための基本語用例辞典 *(第三版)*, 大蔵省印刷局, 1990.