6Q－9

決定木学習アルゴリズムを用いた多義語の訳語選択手法

水野秀紀† 荒木健治†† 宮永喜一† 栃内香次†
†北海道大学大学院工学研究科 ††北海学園大学工学部

1 はじめに

機械翻訳システムの多義語の訳語選択に関して、近年様々な研究が行なわれている [1]。そして、その多くは格情報を用いている。格情報を用いるためには構文解析ツールを用いる必要があるが、その結果には誤りが含まれている。そこで本稿では、格情報を用いずに、距離情報より、事例から決定木学習アルゴリズム [2] を用いて決定木を構築し、それを基に多義語の訳語選択を行なう手法を提案する。

2 システムの概要

本システムは、図1のように学習部、多義性判定部、フィードバック部から構成される。

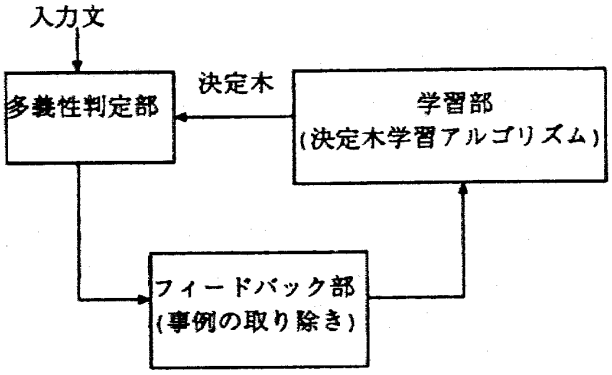

図 1: システムの構成

2.1 学習部

本システムの学習部では、決定木学習アルゴリズムとしてC4.5[2] を用いる。このアルゴリズムは属性、属性値、クラスから成る事例集合が与えられると、情報理論的なヒューリスティック関数を最大にする属性を選択して事例集合を再帰的に分類し、クラスを完全に弁別出来る決定木を構成する手法である。決定木の節点は、1つの属性を調べるためのテストを指す。

本システムでは、1つの事例が、多義性の解消を行なう語を含む文を構成する1単語を表す。そして、それぞれの事例の属性は、多義性の解消を行なう語からの文上の距離、事例における頻度、シソーラスの最上位概念それぞれをとる確率、シソーラスにない品詞(代名詞·前置詞·助動詞·接続詞)になり得るかどうかの有無を指し、クラスは、多義性の解消を行なう語の意味を指す。なお、本手法では、シソーラスとして Cognitive Science Laboratory, Princeton University で開発されたフリーソフト WordNet を用いている。

2.2 多義性判定部

多義性判定部では、学習部で作られた決定木を基に入力文の各単語を分類していく。その際、事例に存在しない単語に関しては、シソーラスから同義語・上位語・下位語を抽出し [3]、抽出した語で分類を行なう。そして、式(1)の値が最も大きいクラスを最適解として選択する。

$$Y_i = \sum_{j=1}^{N} w_j X_{ij} \qquad (1)$$

X_{ij}は分類により求まったそれぞれの語のクラス i の確信度を表し、w_jはそれぞれの語のシソーラス上の距離に応じた重み係数を表す。また、N は入力文の総語数を表す。

2.3 フィードバック部

フィードバック部では、事例を適応的に構築するために事例の取り除き処理を行なう。取り除き方法は、クラスの異なる同じ単語があった場合にその単語同士の距離(多義性の解消を行なう語からの文上の距離)の差を計算し、その逆数をその単語の取り除き指数に加えていく。そし

Method for Word Selection of Multi-Meaning Words Using Decision Tree Learning Algorithm
Hideki Mizuno†, Kenji Araki††,
Yoshikazu Miyanaga† and Koji Tochinai†
†Graduate Shool of Engineering,
Hokkaido University
††Faculty of Engineering,
Hokkai-Gakuen University

表 1: 分析結果

	正解		誤り	
	個数 (個)	割合 (%)	個数 (個)	割合 (%)
既知データ (1)	78	18.7	5	1.2
既知データ (2)	6	1.4	248	59.3
未知データ	25	6.0	56	13.4
合計	109	26.1	309	73.9

て、取り除き指数がある閾値を越えたらその単語は事例から取り除かれる。このことにより、クラスの異なる属性の同じ事例が取り除かれ、決定木の精度が向上すると考えられる。また、全ての事例が取り除かれないようにするために取り除き回数がある一定値を越えた時にその単語を事例に復活させる処理を行なう。

3 実験と考察

3.1 実験

実験は多義性の解消を行なう語を take とし、最初に事例の初期状態を空にしてからC言語のマニュアル[4][5]内の take を含む英文とその日本語訳文 24 組を用いて図 1に示す方法で行なった。

実験の結果、正解率が43.8%となった。また、訓練事例に正解が存在しなかったものが 8 文あった。

3.2 考察

実験で入力したそれぞれの単語の分析結果を表1に示す。表1において、未知データとは分類する際にテストとして選択された属性集合の属性値の組合せが訓練事例になかったものを指し、既知データは逆にその組合せが訓練事例にあったものを指す。そして、既知データ (1) は入力語の正解のクラスと、テストとして選択された属性集合の属性値の組合せが入力語と等しい訓練事例のクラスとが一致したものであり、既知データ (2) は逆に一致しなかったものである。

表 1より、誤りの既知データ (2) が最も多いが、これは入力語とほぼ同じ属性値を持つ訓練事例が存在するがクラスが異なるために誤ったもので、事例不足が原因であると考えられる。

また、今回事例に復活させる処理を行なったが、この処理により属性値が全く同じでありながらクラスが異なる事例が出来、決定木の精度を下げてしまったものと考えられる。

また、今回平均 25 単語からなる 1 文全ての単語を入力したが、予備実験により、take の意味を決定する際に必要な単語数は平均 7 単語であることが分かり、入力単語数を減らす必要があると考えられる。

4 おわりに

本稿では、格情報を用いずに、距離情報などを基に決定木アルゴリズムを用いて多義語の訳語選択を行なう手法について述べ、小規模な実験ではあるが、本手法が有効である可能性を示すことができた。今後本手法の有効性を確かめるためにさらなる実験を行なう予定である。

参考文献

[1] 藤井 敦, 乾 健太郎, 徳永 健伸, 田中 穂積: 動詞多義性解消における格要素の貢献度について, 情報処理学会研究報告, 96-NL-111, pp.55-62 (1996-1).

[2] Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).

[3] 水野 秀紀, 荒木 健治, 宮永 喜一, 栃内 香次: 機械翻訳における事例を用いた多義語の訳語選択手法, 情報処理学会研究報告, 97-NL-120, pp.7-13 (1997-7).

[4] Brian W.Kernighan and Dennis M.Ritchie: The C programming language. Second Edition, Prentice-Hall (1988).

[5] B.W. カーニハン, D.M. リッチー著, 石田晴久訳: プログラム言語C第2版, 共立出版 (1989).