帰納的学習を用いた適応型多言語情報検索手法

1U－1

王 忠建　荒木健治　栃内香次
北海道大学大学院工学研究科

1　はじめに

近年、情報化社会の発展に伴い、電子化テキストデータの普及が著しく、大量の文書を必要に応じて容易かつ効率良く柔軟に検索できる文書検索技術の確立が必要である。しかし、このような大量の文書の検索に関する従来の研究[1][2]は、検索対象が同一の言語である。そのため、最近インターネットの普及により必要となった多言語を対象として十分な性能を発揮する検索システムはまだ開発されていない。ユーザの必要とする検索結果は人によって異なると考えられる。更に入力語を英語キーワードとする時、訳語の多義性のため、検索結果は誤りを多く含んでいる可能性がある。いかに検索結果を絞り、ユーザに適応するかが重要な課題と考えられる。そこで本稿では、ユーザに適応するために、同義語の検索と類似検索を行ない、次にフィードバックされた検索結果を形態素解析し、検索対象文内の情報を利用する。また、本手法は検索結果から帰納的学習[3]を用いて、同じ内容であっても同義語の存在など表現が異なるための検索、逆に、同じ表現でも違う内容の検索結果を絞るなどの方法を用いてユーザに動的に適応する。

2　処理の概要

本手法の処理過程を図1に示す。

2.1　翻訳部

翻訳部では、入力語の英語キーワードを日本語に翻訳し、シソーラス[4]を用いて、同義語を求めて、検索語とする。更に、検索の漏れを軽減するため、類似検索を利用する。

A Method of Cross Language Information Retrieval for Adapting User by Inductive Learning Algorithm
Zhongjian Wang
Kenji Araki
Koji Tochinai
Graduate Shool of Engineering,
Hokkaido University

2.2　検索部

検索部では、訳語を検索語として、学習で獲得した参照情報(正解と誤りのそれぞれの共通単語情報)を参照して全文検索を行なう。

2.3　フィードバック部

フィードバック部では、ユーザは検索結果からランダムに選択したそれぞれ同じ数の正解と誤りを、学習部にフィードバックし、システムに学習させる。

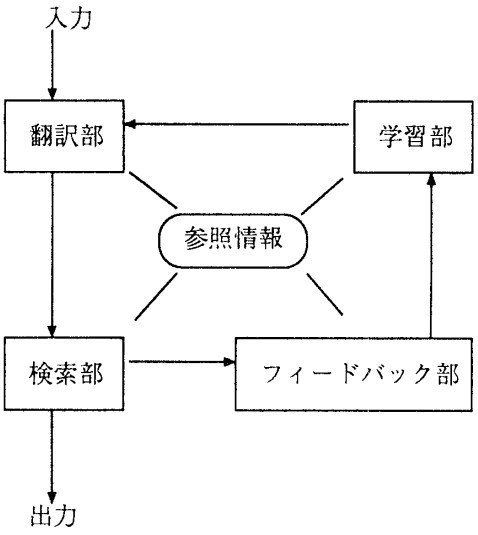

図 1: 処理の流れ

2.4　学習部

ここで、ユーザに適応する性能を高めるために、帰納的学習方法を用いている。フィードバックされた正解と誤りの検索結果を形態素解析し、それぞれの共通単語の出現頻度を計算し、ユーザの適応の検索情報を帰納的学習で抽出する。そして、検索を繰り返し、この適応情報を利用して絞った検索結果の正解と誤りを人手で判断して、フィードバックし、適合率, 式(2)を用いて

表 1: 検索結果

検索質問	フィードバック前				フィードバック後				
	検索結果	正解	誤り	適合率%	検索結果	正解	誤り	適合率%	再現率%
language	66	44	22	63.6	53	44	9	83.0	100
law	139	22	117	18.6	30	22	8	73.3	100
program	145	46	99	31.7	59	43	16	72.8	93.5

自動的に適合率を計算し、それと検索結果の数でゆう度, 式 (1) の中のβ値を調節し、ある閾値を満足する適合率の検索結果の中、最も数が多い検索結果を出力する。

$$\text{ゆう度値}\,(V) = \alpha\frac{\sum(CH)}{\sum(S)} - \beta\frac{\sum(EH)}{\sum(S)} \quad (1)$$

式 (1) において、αとβは重み係数、$\sum(S)$ は一つ検索結果文中の単語の数。正解と誤りのそれぞれの共通単語は検索対象文の長さに依存しないように$\sum(S)$ を用いて正規化する。$\sum(CH)$、$\sum(EH)$ は検索結果の一文中での正解と誤り単語の出現頻度の和を表す。

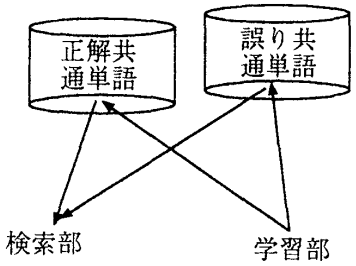

図 2: 参照情報

3 評価実験

実験のデータはインターネットから引用した文書を用いた。英単語を検索語とした全文検索結果を絞るためにゆう度式を利用した。ゆう度式の中に、重み係数αの値とゆう度の閾値を予め固定値とする。そして、適合率によってシステムが自動的にβを調節し、ユーザに適応した検索結果を求める。表1 はフィードバックした前とフィードバックした後の適応した検索結果を示す。

ここで検索結果を評価するのは適合率の式 (2) と再現率の式 (3) を用いている。表 1 からフィードバック後の適合率と再現率がかなり良い数値を示している。

$$\text{適合率} = \frac{\text{正検索結果数}}{\text{全検索結果数}} \times 100\% \quad (2)$$

$$\text{再現率} = \frac{\text{正検索結果数}}{\text{全正検索結果数}} \times 100\% \quad (3)$$

4 おわりに

本稿では、多言語テキストの情報を検索するという立場から、帰納的学習を用い、ユーザに動的に適応する多言語の情報検索手法を提案した。多言語を対象とした情報検索手法は大きな需要があると考えられる。小規模な実験により本稿提案した方法の有効性を示すことができた。今後は、実験の規模を拡大し、評価する予定である。

謝辞

本研究の一部は文部省科学研究費補助金 (課題番号 10680367) により行なわれた。また、以下のツールを使用しました。ここに感謝致します。

京都大学工学部 長尾研究室, 奈良先端科学技術大学院大学 松本研究室:日本語形態素解析システム JUMAN

参考文献

[1] 長尾 真：自然言語処理, 岩波書店 (1996.4)

[2] 河瀬剛, 佐藤理史： 超並列計算機を用いた全文検索の高速化, 情報処理学会第 48 回全国大会論文集, 4E-03,1994.3.

[3] 荒木 健治, 高橋 祐治, 桃内 佳雄, 栃内 香次: 帰納的学習を用いたべた書き文のかな漢字変換, 電子情報通信学会論文誌 D-II Vol.J79-D-II No.3 pp.391-402 1996.3

[4] 国立国語研究所：分類語彙表, 秀英出版社