日本語マニュアルの内容検索システム

4K－4

松崎 知美　三浦 健仁　小俣 祐介　斉藤 貴也　山田 剛一　森辰則　中川裕志*
横浜国立大学 工学部

1　はじめに

電子機器やソフトウェアを使っていて分からないことがあった時、質問文を打ち込めば、マニュアルの読むべき部分を示してくれるシステムがあればユーザの助けとなるであろう。現在このようなときに助けとなるものとして、WINDOWSのオンラインHELP機能などがあげられる。しかし、このようなことができるためには、あらかじめマニュアルライターらが、そのための文章を書いておかねばならない。そのような機能が付与されていないマニュアルでも、質問を打ち込まれれば、すぐに内容を検索して、読むべき部分を示すことのできる、マニュアル内容検索システムが容易に構築できることが切望される。そこで本研究ではこのようなマニュアル内容検索システムを既存のテキスト形式のマニュアルから自動構築する方法に付いて検討する。

2　システムの動作

内容検索システムの構造を図1に示す。システムはユーザからマニュアルに対する質問文を受け付け、その質問の答となるような読むべき部分を表示する。

質問文とマニュアルの両方とも、「の」や名詞からなる名詞句のみを取り出し、検索を行なった。各セグメントのマッチングスコアの計算の仕方は、名詞のtf.idf重みを用いた標準的なベクトル空間法と、複合名詞の最長一致部分のtf.idfの和をとる方法について比較検討を行なった。

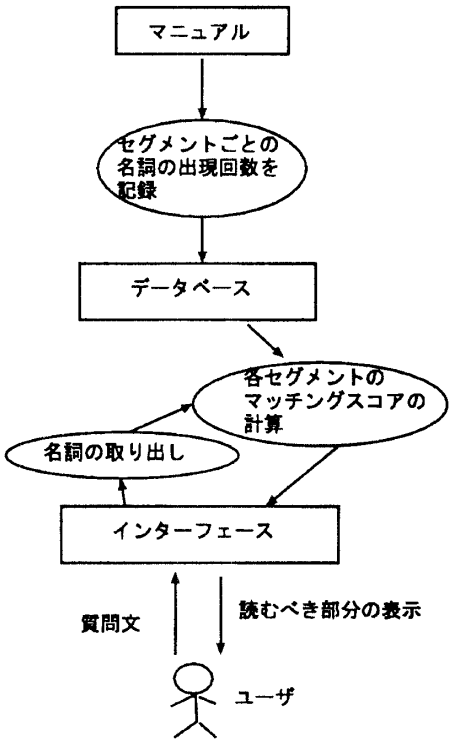

図 1: 内容検索システムの構造

3　評価

実際に表1に示すマニュアルについて、それぞれのマニュアルに対し20問程度の質問を集め、それに対する正解を人手で調べて検索システムの評価を行なった。ここで問題になるのは検索の単位となるセグメントの決め方である。これに関して、1) 章、節のうち最小の形式的構成要素(例えばサブセクションなど)を用いる方法、2) 固定長のセグメント、具体的には10、20、40行の各々の長さの固定長セグメントを用いる方法、について評価実験を行なった。

ここで、質問に対する正解セグメントの決定法としては、あくまで質問に対して答えているものとし、関連があってもそれだけでは質問の答とならないと判断された

表 1: 評価に用いたマニュアルと質問数

マニュアル	size (kB)	質問数
日本語形態素解析システム JUMAN	31	20
構文解析システムSAX	29	24
家庭用ビデオデッキ	69	21
仮名漢字変換フロントエンドプロセッサ「たまご」	57	20

表 2: 検索結果として得られるセグメントの数

マニュアル	ベクトル空間法			複合語tf.idf法		
	min	max	平均	min	max	平均
JUMAN	1	35	22.2	1	35	22.1
SAX	1	23	9.6	1	23	8.2
ビデオ	1	24	15.9	1	24	15.2
たまご	0	43	24.5	0	43	20.5

*Contents Retrieval System of Japanese Manual
by Tomomi Matsuzaki, Takehito Miura, Yusuke Komata, Takaya Saitou, Kouichi Yamada, Tatsunori Mori and Hiroshi Nakagawa,
Yokohama National University, 79-5,Tokiwadai, Hodogaya-ku, Yokohama 240, Japan.

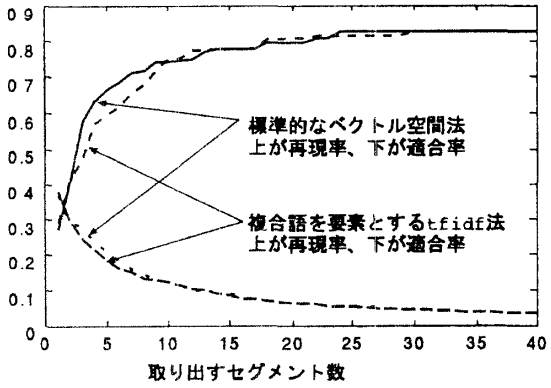

図 2: 最小形式的構成要素をセグメントとする場合で2つの方法でスコア上位のものから取り出した際の適合率・再現率

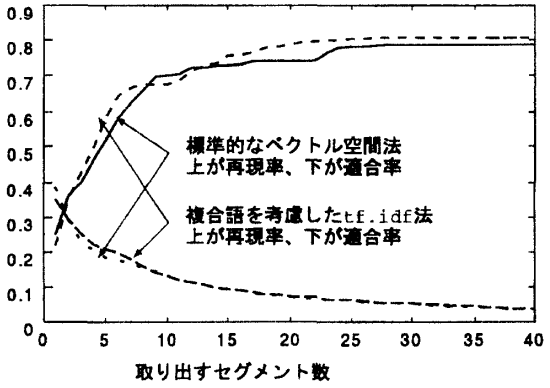

図 3: 固定長セグメント (20 行) を用いた際の適合率・再現率

ものは不正解とした。このため1問の質問に対する正解とされたセグメントの数は最も多いもので7セグメントあった。一方、マニュアル中に適当な正解が無いと判断された質問もあった。このような質問は再現率、適合率ともに0として扱った。また、一つの質問に対して、システムが検索結果として返すセグメントの数は表2の通りであった。(検索結果としてはスコアが0より大きいセグメントが全て返される。)

1) 最小の形式的構成要素をセグメントとする場合

標準的なベクトル空間法と、複合語を要素とするtf.idf法によるスコアでセグメントをランキングして上位から取り出していった際の適合率、再現率を調べて評価を行なった。4つのマニュアル、計85問の質問について実験した。それぞれについてスコア1位のセグメントを取り出した際の適合率・再現率、スコア2位までのセグメントを取り出した際の適合率・再現率、…、スコアn位までのセグメントを取り出した際の適合率・再現率を求め、質問1問あたりの平均値を出した。結果を図2に示す。ランキング1位のセグメントの適合率・再現率は、ベクトル空間法で37.5%・27.4%、複合語tf.idf法で38.6%・29.3%であり、複合語tf.idf法が勝っている。しかし、グラフの再現率の2本の曲線を見ると、ベクトル空間法の方が取り出すセグメント数を増やした際に再現率が急激に上昇することが分かる。実際の値で見ても取り出すセグメント数を3セグメント以上にするとベクトル空間法が勝る。

2) 固定長セグメントの場合

ここではマニュアルを10行ごと、20行ごと、40行ごとに、区切った際の適合率、再現率を調べた結果を述べる。ベクトル空間法で、マニュアルを10行ごとに区切った際で、ランキング1位のものを取り出した適合率29.1%、再現率14.6%、同様にマニュアルを20行ごとに区切った際で、適合率38.7%、再現率24.9%、40行ごとに区切った際で適合率36.4%、再現率26.1%であった。同様にして複合語を考慮したtf.idf法で、マニュアルを10行ごとに区切ると適合率34.1%、再現率17.2%、20行ごとに区切ると適合率35.3%、再現率21.9%、40行ごとに区切ると適合率37.5%、再現率25.2%であった。図3はマニュアルを20行ごとに区切ってセグメントとした際の適合率・再現率である。章・節を利用した場合に比べて5%程度、再現率が劣っている。

この両方法を比較検討してみる。セグメントを小さく定めると、結果が正しければ、質問に対する答をより局所的に限定できて分かりやすいが、それだけ検索の精度をあげるのが難しい。マニュアルのライターが定めた章や節には意味的なまとまりがあることが予想され、検索結果としても読みやすい。しかし、場合によってはセグメントが長いものとなってしまい、セグメントを指定されただけでは答を述べている部分を探すのに手間がかかってしまう場合がある。このような問題はセグメント表示インターフェースを工夫して解決することを試みた。以下にこれについて述べる。

4 セグメント表示インターフェース

HTMLで書かれたマニュアルに対しHTMLブラウザを用いてユーザからの質問文入力と検索結果の表示を行なった。検索結果としてランキングされたセグメント番号を得る。その際にセグメントの先頭からではなく、そのセグメントではじめて出現する質問文中の名詞および名詞句の周辺部分から表示を行なう。これは質問文中の名詞が存在している文がユーザの目的の部分であると仮定して、質問文中の名詞がない文は目的の部分ではないと判断し、読む必要のない部分を省いて目的の部分から表示するためである。質問文中の名詞および名詞句を目立つように太文字表示してユーザの利便を図っている。

5 おわりに

表示インターフェースの改善と適合率、再現率の改善が今後の課題である。

謝辞

本研究はIPAの創造的ソフトウェアプロジェクトの援助を受けている。