日本語マニュアルの内容検索システム

三浦 健仁　松崎 知美　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

最近の電子機器、ソフトウェアなどのマニュアルは決して読みやすいものではない。質問に応じて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。既存のものでは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多大な手間をかけて、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、質問に応じてすぐに内容を検索し、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を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　システムの構造

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。

まず、対象マニュアルについて、以下の作業を行ない、検索のためのデータベースを用意する。

1. マニュアルを章、節といった構造を元にするなどの方法でセグメント単位に分割する。
2. マニュアルの文を形態素解析システムJUMANを用いて単語に分割する。
3. 名詞および複合名詞と名詞どうしが「の」で連接された名詞句を全て取り出す。
4. セグメントごとの名詞の出現回数を求める。
5. tf.idfによる単語の重みの計算をしておく。

このようなデータベースが実際に用意してある状態で、ユーザが質問文を打ち込むと次のような手順の末、結果を表示する。

1. 質問文を形態素解析システムJUMANにかける。
2. 名詞および複合名詞と名詞どうしが「の」で連接された名詞句を全て取り出す。
3. データベースを用いて質問文から取り出した名詞句と各セグメントのマッチングスコアを計算する。
4. 各セグメントをマッチングスコアの高い順にランキングする。
5. ランキング順にセグメントを表示する。

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。

3　スコアの計算

3.1　標準的な名詞のtf.idfを用いたベクトル空間法

これは広く一般的に行なわれている方法である。まず個々のセグメントに出てくる名詞についてのベクトルを求める。ただし、名詞に与える重みはtf.idf重みとする。名詞NのセグメントS_jにおける重み$w(S_j, N)$の定義式は以下の通りである。

$$w(S_j, N) = tf(S_j, N) \times idf(N) \tag{1}$$

$$idf(N) = \log_2\left(\frac{\#Seg}{freq(N)}\right) + 1 \tag{2}$$

$tf(S_j, N)$：セグメントS_jにおける名詞Nの出現回数
$\#Seg$：マニュアルを分割したセグメント数
$freq(N)$：名詞Nが出現するセグメントの数
各セグメントで出現する全ての名詞についてこの値を求めるが、複合語は構成する名詞の最小単位に分割して扱う。あるセグメントS_jに対するベクトルは、セグメントに現れた名詞と質問文に現れた名詞が併せてm種類、名詞$N_1, N_2, \cdots, N_m$のとき、次の式で表される。

$$\overrightarrow{S_j} = (w(S_j, N_1), w(S_j, N_2), \cdots, w(S_j, N_m)) \tag{3}$$

質問文のベクトルもセグメントの場合と同様に定義する。先ほどと同様、セグメントに現れた名詞と質問文に現れた名詞が併せてm種類、名詞$N_1, N_2, \cdots, N_m$であるとすると、質問Qに与えるベクトルは次の式で表される。

$$\overrightarrow{Q} = (a(N_1), a(N_2), \cdots, a(N_m)) \tag{4}$$

$a(N)$：名詞Nが質問中に現れれば1、現れなければ0
この両者のベクトルの*cosine*で、質問文QとセグメントS_jの類似度が求められる。

$$Similarity(S_j, Q) = \frac{\overrightarrow{S_j} \cdot \overrightarrow{Q}}{|\overrightarrow{S_j}| \cdot |\overrightarrow{Q}|} \tag{5}$$

3.2　複合語のマッチングを考慮したtf.idf法

前節のベクトル空間法では名詞は単名詞に分割して計算を行ない、複合語どうしの一致については考慮しなかったが、質問文の分解結果にも検索対象のマニュ

アル本文中にも複合語が現れる。複合語どうしが完全に一致する事もあるが、片方が他方の複合語の一部となっていたり、部分的に一致する場合もある。このような場合に独立の名詞としてではなく、連続した名詞の一致として扱う方法について述べる。

そこでまず、質問文の分解結果のうちの一つの名詞句と、検索対象のマニュアル中の一つの名詞句のマッチングに注目する。質問文中の名詞句 $W_k^{\mathbf{Q}}(k = 1,2,\cdots,N)$ の集合を $\mathbf{Q}$、セグメントに分割したマニュアルの j 番めのセグメント S_j 中の名詞句 $W_h^{\mathbf{S}_j}(h = 1,2,\cdots,M)$ の集合を $\mathbf{S}_j$ とおく。

$$\mathbf{Q} = \{W_k^{\mathbf{Q}} | k = 1,2,\cdots,N\} \quad (6)$$

$$\mathbf{S}_j = \{W_h^{\mathbf{S}_j} | h = 1,2,\cdots,M\} \quad (7)$$

$W_k^{\mathbf{Q}}, W_h^{\mathbf{S}_j}$ のそれぞれが名詞 N_i の並びからなる名詞句であるとする。

$$W_k^{\mathbf{Q}} = /N_1/N_2/\cdots/N_n/ \quad (8)$$

$$W_h^{\mathbf{S}_j} = /N_1/N_2/\cdots/N_m/ \quad (9)$$

この2つの名詞句から一致している部分を取り出す。そのさい連続して一致している部分はひとまとまりとして取り出す。例えば次のような場合 (/ は名詞どうしの連接を表す) について考える。

$$W_k^{\mathbf{Q}} = /A/B/C/D/E/ \quad (10)$$

$$W_h^{\mathbf{S}_j} = /B/C/E/ \quad (11)$$

この場合、取り出される部分は $/B/C/$ と $/E/$ である。

なお、ひとつの複合語中に同じ名詞が2回以上出現する場合には構成する名詞の数が多いパターンを優先して取り出す事にする。例えば、

$$W_k^{\mathbf{Q}} = /A/B/D/B/C/E/ \quad (12)$$

$$W_h^{\mathbf{S}_j} = /A/B/C/E/ \quad (13)$$

のような場合、$/A/B/$ と $/B/C/E/$ というパターンが考えられる。このような場合は $/B/C/E/$ を取り出してから、残りの $/A/B/D/$ と $/A/$ を比較し $/A/$ を取り出す。

さて、取り出したパターンをそれぞれ

$$P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_1, P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_2, \cdots, P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_r$$

とする。上の例では、

$$P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_1 = /B/C/E/, P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_2 = /A/$$

となる。ここで、

$$pat(W_k^{\mathbf{Q}}, W_h^{\mathbf{S}_j}) = \{P(W_h^{\mathbf{S}_j}, W_k^{\mathbf{Q}})_i | i = 1,\cdots,r\} \quad (14)$$

としたとき、セグメント S_j と質問中の名詞句 $W_k^{\mathbf{Q}}$ から得たパターンの集合を次のように表す。

$$\{P(S_j, W_k^{\mathbf{Q}})_i | i = 1,\cdots,R\} = \bigcup_h pat(W_k^{\mathbf{Q}}, W_h^{S_j}) \quad (15)$$

そして、$P(S_j, W_k^{\mathbf{Q}})_i = /N_1/N_2/\cdots/N_l/$ というパターンに次のような重みを与える。

$$pw(S_j, P(S_j, W_k^{\mathbf{Q}})_i) = tf(S_j, P(S_j, W_k^{\mathbf{Q}})_i) \times idf(P(S_j, W_k^{\mathbf{Q}})_i) \quad (16)$$

パターンのスコアを求める際に、パターンの tf.idf を用いている。tf.idf はパターン全体について求めている。tf はスコアを求めるセグメント内でのパターン $P(S_j, W_k^{\mathbf{Q}})_i$ の出現回数である。idf の計算式は次式を用いた。

$$idf(P(S_j, W_k^{\mathbf{Q}})_i) = \left(\log_2 \frac{\#Seg}{freq(P(S_j, W_k^{\mathbf{Q}})_i)}\right) + 1 \quad (17)$$

$\#Seg$：一つのマニュアルが分割されたセグメントの数
$freq(P(S_j, W_k^{\mathbf{Q}})_i)$：一つのマニュアル中でパターン $P(S_j, W_k^{\mathbf{Q}})_i$ が出現するセグメントの数

質問中の一語、$W_k^{\mathbf{Q}}$ に対するスコアを次の式で与える。

$$WScore(S_j, W_k^{\mathbf{Q}}) = \sum_{l=1}^{R} pw(S_j, P(S_j, W_k^{\mathbf{Q}})_l) \quad (18)$$

質問文に対するセグメントのスコアは、質問文に出現した全ての名詞句 (単名詞および複合名詞) に対するスコアを合計し、セグメントに対して次式のように正規化したものとする。

$$SScore(S_j, \mathbf{Q}) = \frac{\sum_{k=1}^{N} WScore(S_j, W_k^{\mathbf{Q}})}{\sqrt{\sum_{l=1}^{M} (pw(S_j, W_h^{\mathbf{S}_j}))^2}} \quad (19)$$

3.3 タイトル語や共起する語の重みづけ

検索部の計算方法を修正し適合率、再現率の改善をはかった。

3.3.1 タイトルの扱いについて

セグメント内にタイトルが含まれる場合、タイトルは書かれている内容を示す重大な情報となるだろう。そこでタイトルに出てくる名詞に重みを与えるという実験を行なった。

標準的なベクトル空間法へのタイトル語の重みづけ ある名詞の tf.idf 重みを計算する際に、その名詞がそのセクションでタイトルに出現する名詞だった場合、その名詞のそのセクションにおける tf.idf 重みを 1.5 から 3 倍した。

複合語のマッチングを考慮した tf.idf 法へのタイトル語の重みづけ 質問文とマニュアルから取り出された一致パターンの tf.idf を計算する際に、この一致パターンがあるセグメントにおいてタイトルに現れたパターンだった場合、そのセグメントにおける tf.idf の値を 1.5 から 3 倍にすることにより重みづけを行なった。

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 (個)
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

3.3.2 共起する語の扱いについて

セグメント中で文内共起する語が共に検索された場合、そのセグメントのスコアを 1.5 から 3 倍するという重みづけを行なった。これは名詞を全て個々に扱うベクトル空間法と複合語を考慮した tf.idf 法で重みづけの方法が異る。

標準的なベクトル空間法への共起があった際の重みづけ 検索された 2 個以上の名詞が、あるセグメントで同一の文中に現れていた場合、そのセグメントのスコアを共起するのペア数回だけ 1.5 から 3 倍する。これを行なうことにより結果的に検索された複合語とマニュアル中の複合語が一致した場合のスコアが上がる。

複合語のマッチングを考慮した tf.idf 法への共起があった際の重みづけ あるセグメントで取り出された一致パターン 2 つが、そのセグメントにおいて同一の文中に現れた名詞句 2 つと完全一致した場合、そのセグメントのスコアを 1.5 から 3 倍した。

4 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 (そのうち 50% をオーバーラップするものを含む) を用いる方法、の 2 つ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

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

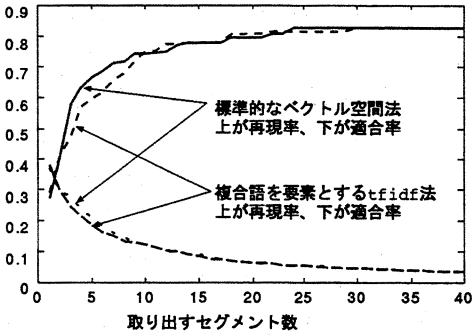

図 1: 最小形式的構成要素をセグメントとする場合で 2 つの方法でスコア上位のものから取り出した際の適合率・再現率

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。

4.1 最小の形式的構成要素をセグメントとする場合

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・再現率からスコア 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 1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。

4.2 固定長セグメントの場合

マニュアルを 10 行ごと、20 行ごと、40 行ごとに、区切った際のランキング 1 位の適合率、再現率を表 3 に示す。 図 2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。

両方法を比較すると、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。逆に章や節には意味的なまとまりがあるこ

表 3: 固定長セグメントでランキング 1 位を取り出した際の適合率・再現率

セグメント長 (行)	ベクトル空間法		複合語 tf.idf 法	
	再現率	適合率	再現率	適合率
10	14.6	29.1	17.2	34.1
	(8.4)	(35.4)	(7.7)	(34.3)
20	24.9	38.7	21.9	35.3
	(13.2)	(41.3)	(13.0)	(41.2)
40	26.1	36.4	25.2	37.5
	(18.5)	(50.6)	(15.4)	(45.9)

() 内は 50% オーバーラップした場合

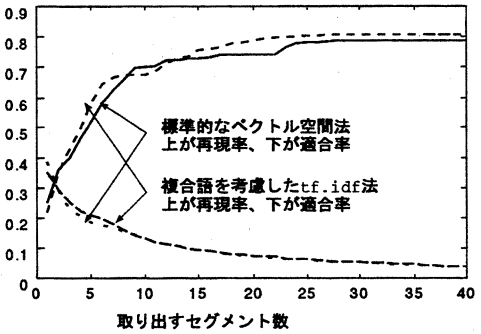

図 2: 固定長セグメント (20 行) を用いた際の適合率・再現率

とが予想され検索結果としても読みやすいが、場合によってはセグメントが長くなってしまい指定されたセグメント内で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フェースを工夫して解決することを試みた。

5 セグメント表示インタフェース

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て、その内容を表示する。その際に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。

例として質問文に「メールにファイルを添付するには」を入力した場合の表示結果を図 3に示す。

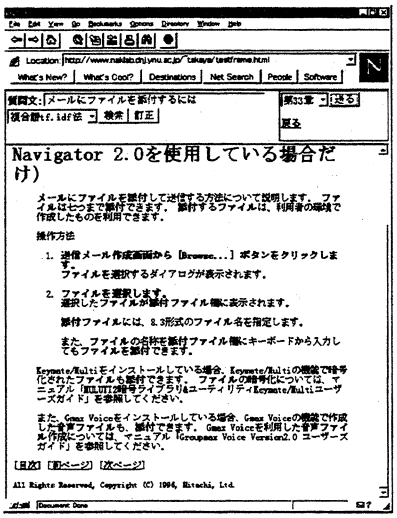

図 3: インタフェース画面 (日立製作所より提供頂いたマニュアル「Groupmax World Wide Web ユーザーズガイド」より)

6 おわりに

本稿ではマニュアル内容検索システムを提案し、内容検索の評価実験を行なった。表示インタ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。

参考文献

[WAIS 93] WAIS : WAIS Server, WAIS Workstation, WAIS Forwarder for UNIX, WAIS Inc., Technical Description Release 1.1 (1993).

[Callan 94] Callan, J. P. : Passage_Level Evidence in Document Retrieval, Proc. 17th ACM SIGIR, pp.302-310(1994).

[Wilkinson 94] Wilkinson, R. : Effectibe Retrieval of Structured Documents, Proc. 17th ACM SIGIR, pp.311-317 (1994).

[黒橋 96] 黒橋 禎夫, 白木 伸征, 長尾 眞: 出現密度分布を用いた語の重要説明箇所の特定, 情報処理学会 NL 研究会 -115-7, pp.43-50(1996).

[Salton 93] Gerard Salton, J. Allan and Chris Buckley: Approaches to Passage Retrieval in Full Text Information Systems, 16th ACM SIGIR, pp.49-58(1993).

[Fuller 93] Michael Fuller, Eric Mackie, Ron Sacks-Davis, Ross Wilkinson: Structured Ansers for a Large Structured Document Collection, 16th ACM SIGIR, pp.204-213(1993)