社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－167（11）
2005－SLP－56（11）
2005／5／27

良くある質問（ＦＡＱ）のコンテンツ・統語検索を組み合わせた自然言語質疑応答システム

陳亮　University of Northern British Columbia、
徳田尚之、侯平魁　サン・フレア研究・開発センター
永井明　宇都宮大学総合情報処理センター
陳若愚　瑞安電視大学
鄭然　University of Northern British Columbia

要旨

WORDNET のような開発に膨大な時間の掛かる同義語辞書に頼らずに、素人でも使える大型ＦＡＱシステム向け自然言語の質疑応答システムを構築した。
FAQ システムを自然言語クエリにより検索する仕組みとして、本システムは差分 LSI(DLSI)法によるコンテンツ検索と、テンプレート・オートマトン・マッチングによる統語検索を組み合わせというユニークな構想を持ち、次の３段階処理がその基礎となる。第一段階では、ＦＡＱアイテムとクエリのターム展開の有効性に不可欠な tf-idf 展開を保証するために、質問部と回答部間に存在する語彙ギャップを埋める処理をしたこと、第二段階では、類似意味のコンテンツ検索により、差分 LSI 法を意味的な一次フィルターとして用い、意味的に等価な質問部・回答部ペアのみに絞り込んだこと、第３段階目では、柔軟性に富む自然言語の統語検索機能を持つテンプレートマッチングにより最終的な意味的に等価な表現をもつ、最適なＦＡＱの質問部・回答部ペアに絞り込みユーザに提示する。BURKE 達の編集するのに膨大な労力を必要とする同義語辞書 WORDNET に頼る FAQ Finder.に較べると、我々が開発したこのスキームは簡単に実現可能であり、大幅に労力が削減される。この方法の有効性は Lucene FAQ System の FAQ システムで実験的に実証した。

Combined Content and Syntactic Search-Based QA System in a FAQ Domain

Liang Chen　University of Northern British Columbia、
Naoyuki Tokuda、Pingkui Hou　Sunflare Company R&D Center
Akira Nagai　Utsunomiya University
Ruoyu Chen　Ruian Television University
Ran Zheng　University of Northern British Columbia

Abstract

To facilitate and enhance the usability of a large FAQ system, we have developed a new user-friendly, combined content and syntactic search-based QA system that accepts free format natural language queries from users, and guide them to an answer item by locating and pinpointing the appropriate Q&A items within the FAQ dataset. To maximize the combined effects of both content and syntactic searches, we have introduced a three step core procedure comprising the *term expansion of FAQ items and queries* to ensure the validity of the tf-idf expansion, *the DLSI (differential latent semantic indexing)-based semantic filtering step* to capture semantically similar expressions in content; and *the final flexible and powerful lexical template matching step* to accommodate a rich variety of natural language queries of semantically similar expressions. Unlike the FAQ Finder of (Burke *et al.*, 1997) which depends on a thesaurus-type dictionary such as Word-Net to match question and answer, the new scheme is simple to implement. An experimental investigation using the Lucene FAQ collection confirms the effectiveness of the method.

１ いとぐち

大企業、教育機関、官公庁などの大型のポータルやウェブサイトを外部の人に使いやすくするために頻繁に使われる質問―回答ペアのデータセットからなる“よくある質問”（ＦＡＱ）も、小型にとどまっている範囲では簡単に疑問に対応する回答群を見つけられるが、大型になると何処の質問事項を探せばよいのかが分からず、極端に使い勝手が悪くなる。これを人手で対応していたのでは、ウェブサイトの管理者に極端に負担が増える。自然言語で記述されているドキュメント群の中から、自然言語で記述されたクエリ項目のキーワードを含む文章群を選んで提示するいわゆる TREC(Text Retrieval Conference;http://trec.nist.gov/)方式の質疑応答(QA)システムとの違いは、本ＦＡＱシステムではクエリ項目の検索対象はＦＡＱデータセットの質問部・回答部ペアの両方に現れる可能性があることである。ＦＡＱのＱＡシステムについての研究がほとんど見あたらないのは、これから述べるようにＦＡＱシステムには語彙ギャップという独自の難しさがあるからだが、唯一の FAQ システムによる QA システムの研究は Burke 等(1997)による WORDNET を意味知識データベースとして使った FAQ Finder であるが、実用システムとして使うと長期間をかけて完成しているはずの WORDNET でも頻繁に対応する同義語を見つけられない問題が起きたという。本研究は WORDNET に頼らない新しい仕組みのＦＡＱドメインのＱＡシステムを開発する事を目標に開始された。

情報検索技術で最も広く頻繁に使われる手段は、単語/用語などの文字レベルのマッチング検索手段という表層レベルの統語に基づくマッチングによるもので、現在使われている殆どの検索システムはこの方法による。この方法の一番の問題点は、自然言語の持つ曖昧性からくる用語・単語の同義語の多さ、用語・単語の多義性、修飾句の係り合いの曖昧性などからくる不安定性からくる。この曖昧性がドキュメント全体さらにはその単語の出現する前後の文章に依存するという文脈依存性に起因する。実際もっと範囲の広い世界知識とも呼ばれるいわゆる“常識”に依存する場合も多く、人間の脳の持つ言語創作にまつわる深遠な多くの未開な知識機能の解明なしには解

決できない問題と思われる。これまで以上の統語検索効率を上げるには、WORDNET のような同義語辞書の完備、さらに理想的には Grueber(1995)の唱える「読者・文脈に依存しない仕組みで意志疎通を可能にする用語の概念化プロセスの本格的な定義を提供する」オントロジの発展が望まれるところである。

自然言語の持つ同意語・多義語の多さを統一する仕組みとしては、似た意味の用語を選択するコンテンツ検索の重要性が認識された。もしそれが可能であれば、多言語環境での言語処理に役立つ。コンテンツ検索で広く使われている手法は、tf-idf と呼ばれるドキュメント中に出現する特徴的単語の抜き出し法すなわち「あるドキュメント内での出現頻度が高い(tf)」単語のうち「他のドキュメントにはあまり出現しないもの(idf)」を「そのドキュメントに特徴的な単語」と考える手法で、　ＬＳＩ法(Latent Semantic Indexing)[Zobel and Moffat 1998],差分 LSI 法(Differential Latent Semantic Indexing)[Chen et.al 2001,Chen et.al 2003]、PCA 法(Principlal Component Analysis)等の手法が開発されている。これらの手法はいずれも膨大なベクトル空間の次元数の低減法である。FAQ 分野での本手法の一番大きな問題点は、ペアを組む質問文と回答文の間に語彙断層(lexical gap)[Berger et.al 2000]が存在し tf-idf 法の有効性が失われることである。それは質問文には 5W-1H 等の疑問文系の疑問代名詞が多発するが、回答文には滅多に現れないことからも解る。この語彙の確率分布の均等化を図る方法については次章で述べる。テキスト検索としてのコンテンツ検索法の大きな特徴は、自然言語の本質的な難問、文脈依存性のため生じる同義語、多義語間の曖昧性、さらには係り合いの曖昧性の解消を、編纂などに多大な手間のかかる類似語辞典・オントロジに直結する用語辞書に頼ることなく解決出来ることにあると言って良い。[Burke et.al 1997]の実験でも指摘されている WORDNET の不完全性を考えるともしこのコンテンツ検索法が利用出来れば手間のかかる編纂など類似語辞典・オントロジに直結する用語辞書に頼ることなく解決出来ることになり大きなメリットが生まれる。LSI 法に較べて情報量の多い DLSI 法では、当該ベクトルの持つ LSI 法で使う低減基本ベクトル空間との角度の他にそのベクトル空間までの距離をも使うことによって検索効率を高めている。ＤＬＳＩ法によるコンテンツ検索は、言語に依存しないクロス言語性を使って同義的意味をキャプチャすることにより意味的な一時フィルターの役割を果たすと考えて良い。

一方最も広く使われる統語検索は、上の一次的な意味フィルターで集めた検索候補の中からＦＡＱで使われている単語・用語を正確に検索して最終的な絞り込むのに用いる。ただこの統語検索は全く表面的な文字マッチングであるので自然言語の持つ曖昧性のために生じる言語の柔軟性、例えばほぼ同じ意味でも違う表現とか異なる文構造､多様な同義語等による曖昧性の問題は残る。このコンテンツ検索・統語検索でカバー出来ない自然言語の問い合わせシステムを実現するために、意味的に同じ文章を違った表現またはパターンで表現出来る仕組みとして英作文自動添削システム用に開発したテンプレート構造を持ち込み、自然言語によるＦＡＱの検索システムを開発したのでその詳細を発表する。本手法の骨子は次の通りである。

(1) ＦＡＱシステムの質問・回答部分の用語出現頻度の語彙断層を除去する。
(2) 多言語環境で自然言語クエリによるＦＡＱ検索をするためにＦＡＱの質問・回答部間に存在する語彙断層　を除去し、クエリと同等な意味を持つＦＡＱペアを選び出す

(3) 自然言語クエリは、テンプレート構造で納めておき、クエリとは最長共通文字アルゴリズムでマッチングを行う。自然言語による検索を可能にするために、使われる質問に応じてテンプレートは常に更新して最新のものに整備する。

(4) 良くある質問（ＦＡＱ）では必ずしも、検索効率として最も頻繁に使われる Recall, Precision は余り意味を持たない。そのＦＡＱによりどういった評価システムが必要か吟味することが必要である。

２ 質問/クエリ部と回答部の展開

２.１ 質問・回答部の語彙断層の除去法

p(a/q)を、用語ｑが質問部に現れたときに回答部に用語ａが現れる確率、n(a,q)を全 FAQ データセット中で用語 a が回答部に、用語 q が質問部に現れる出現頻度、n(q)を質問部に q が現れるＦＡＱデータセット中の全数とすると

$$p(a|q) = \frac{n(a,q)}{n(q)} \qquad (1)$$

質問$q_1,q_2,\ldots,q_k$が現れたときに、回答部にaが現れたる確率は

$$p(a|q_1,q_2,\ldots,q_k) = 1 - \prod_{i=1}^{k}(1 - p(a|q_i)) \qquad (2)$$

で表される。式２を使うと、質問項目が与えられた時に回答部に現れる項目の確率を計算出来る。逆に回答が与えられれば、質問部に新たな項目・用語を追加して質問・回答部の用語の出現頻度を同等にすることが可能である。自然言語のクエリについても全く同じことが言える。

３　　差分ＬＳＩ(Differential Latent Semantic Indexing)法の適用

“a”,”the”などのストップワードを捨て、FAQ データセットに少なくとも２回は出現する単語、句をタームと定義し、そのタームリストを$t_1,t_2,t_3,\ldots,t_m$と書く。今、質問部のタームリストをＳとし、上記式２を介してリストＳに補足すべき可能性のある用語を抜き出し、この用語リストを補強して新しい用語集を求める過程を質問部の展開という。逆に回答部展開とは、上の展開を回答部について行えばよい。この結果タームベクターとして二組の

質問ベクター　　$(q_1,q_2,q_3,\ldots,q_m)$、

回答ベクター　　$(a_1,a_2,a_3,\ldots a_m)$

を得る。

但し

$q_i=f(q)_i.g_i$、 $a_i=f(a)_i.g_i$

ここで　f(q)i と　f(a)i はそれぞれ質問部および回答部のタームt_iの局所（ローカルな）重み係数を指し、g_iは全てのタームに共通なグローバル重み係数を指す。これらの重み係数は、FAQデータセットの内どのくらいの重要度をそのターム(用語)が持つかを示すパラメータと考えて良い。

用語のグローバルな重み係数の選び方はいくつもあり、生の出現頻度でも、または出現頻度の対数でも良いし、式１で表される確率をとっても良い。良くある質問FAQでは、対応する質問・回答部は一組であり、我々が問い合わせる自然言語のクエリの与え方によっては、一個だけユニークに決まる場合もあれば、全く対応するペアが存在しなかったり、複数個のペアが該当することもあり得る。従ってFAQの検索手続きとしては、質問部と回答部は全く同等と考えて良い。

ここでは、グローバル射影法によるLSI法を、各ドキュメントの特性をより活かした改良版DLSI法(Chen,Tokuda and Nagai、 2001 and、 2003)を適用したロバストな結果について報告する。I_1 とI_2を正規化したタームベクトルとして差分タームベクターをI1,I2 とすると　I=I1-I2 差分正規化ベクトルという。もし、I_1とI_2が同じドキュメントからのタームリストであれば、ベクトルIは内部差分ベクトル、もしI_1 とI_2が違うドキュメントからのベクトルの場合には外部差分ベクトルと呼ぶ。

差分タームドキュメント行列の m 行 n 列でランク r<q=min(m,n)のどんな行列でも U,S,V という三つの行列の積に分解出来る。

$$D = USV^T$$

但し、行列U,Vはそれぞれ　m行q列、q行m列の行列であり、Sは行列DD^T の固有値でq 列 q行のいわゆる降順にソートされた対角行列である。k<rの低次元のU_k, V_k行列をU,Vから求めるにはそれぞれの一番左側k列部分のみを保持すればよい。U_k, S_k とV_k^T
の積は絶対値がほぼDと等しい行列D_kを与える。文献(Chen et.al 2001)が示すように充度関数P(x｜D)は、次式で表される。

$$\hat{P}(x|D) = \frac{n^{1/2}\exp\left(-\frac{n}{2}\sum_{i=1}^{k}\frac{y_i^2}{\delta_i^2}\right)\cdot\exp\left(-\frac{n\epsilon^2(x)}{2\rho}\right)}{(2\pi)^{n/2}\prod_{i=1}^{k}\delta_i\cdot\rho^{(r-k)/2}}, \quad (3)$$

ここで$\underset{\sim}{y} = \underset{\sim}{U}_k^T x = (y_1, y_2, \cdots, y_k)^T$、 $\epsilon^2(x) = \|x\|^2 - \sum_{i=1}^{k} y_i^2$、 $\rho = \frac{1}{r-k}\sum_{i=k+1}^{r}\delta_i^2$　r は行列Ｄのランクである。

実用的には、$\rho = \frac{1}{r-k}\sum_{i=k+1}^{r}\delta_i^2$であり近似値としては$\delta_{k+1}^2/2$として良く、ｒは n で近似して良い。式３での x は、q 及び l を、それぞれ入力クエリと統合したＱ＆Ａ項目の正規化タームベクターとして　x=q - l で表される。マッチングプロセスの詳細は５章で述べる。

４　　FAQ データセットの質問部分項目のテンプレート構造

コンテンツベースの DLSI 法によるサーチ空間のフィルターリングは一次段階での FAQ 検索の意味的な相似空間の絞り込みには有効であるが、ユーザーの入力に沿った正確な検索という意味では統語検索が必要でその問題が残る。FAQ データセットで使う質問文は殆どの場合簡潔な文であり、我々が英作文の語学教育システム（Tokuda and Chen,2001,2003）で開発したテンプレート構造を用いることとした。この仕組みはテンプレートオートマトンと呼ばれ、意味的に同等な文章を組み合わせるのに最も優れている方法であり、ＤＰ法を使った入力文との照合により最速

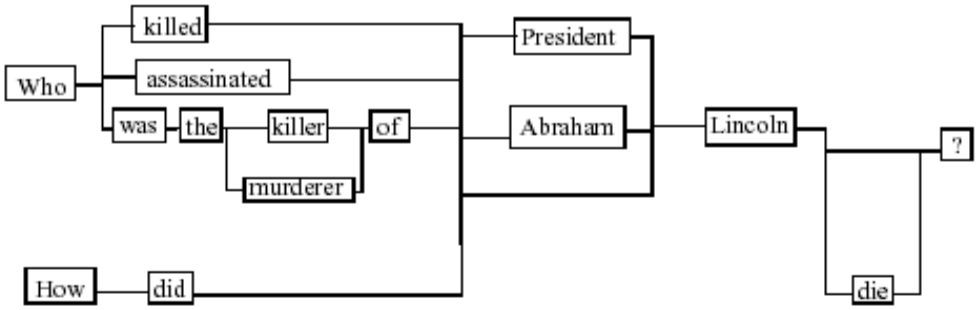

Figure 1: Template Example Indicating a set of Semantically Similar Questions for FAQ

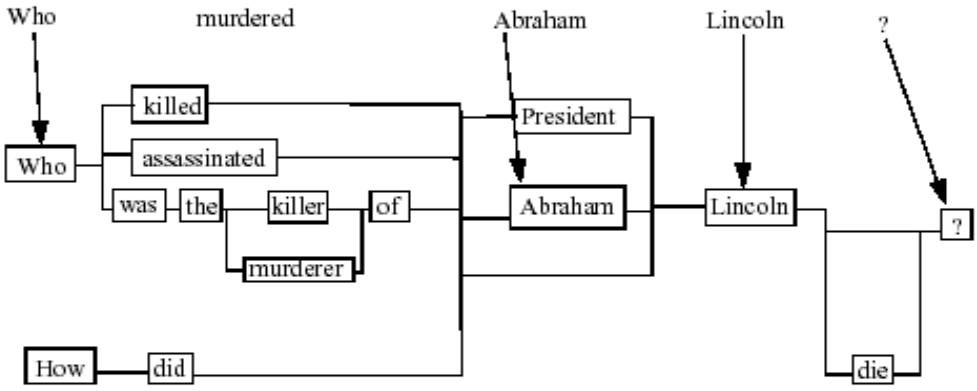

Figure 2: Use's input to be matched with Question Template

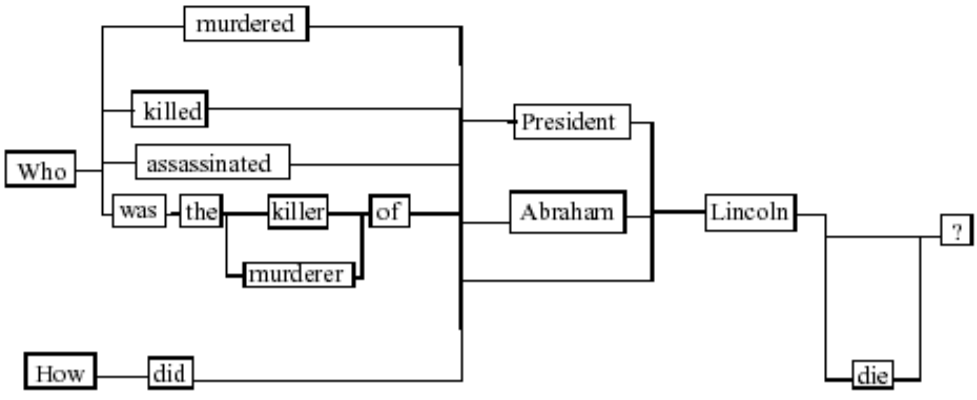

Figure 3: Modified Template

な効率での検索が可能である。

“Who killed Abraham Lincoln?”という簡単な英文入力を例に考察してみよう。図１の DAG(有向非環式グラフ)を使うと例えば”Who was the murderer of President Lincoln?”のような実に多くの意味的に同意な文章表現が可能になる。実際、図１で表現される多くの文例が FAQ データセ

ットの質問・回答群を探し出す質問例として使うことが出来る。例えば “How did Lincoln die?” という質問に対して厳密に言えば、図１の質問文章パーターンと全く同じ質問・回答群を指示する必要性はないが、文章の類似性から同じ質問・回答群を検索しても間違いではないと考えて良い。ではどうやって意味的に同意な文（パス）をテンプレート上に埋め込めるのか？まず、最初のクエリ文をテンプレートに落とし、その後使われるクエリ文を出来たら同じノードに追加するか、もしマッチしなければ新しいノードを加えることが必要である。その過程は図２、図３に示してある。例えば”murdered”という単語は新しいノードに追加されていることに注意願いたい。一見、どのような文でもテンプレート構造を使うと一つのテンプレートに組み込むことが出来そうである.

例えば、”When did Abraham Lincoln die?”を考えてみると、”HOW”のノードに”WHO”を使うと一つのテンプレートに収まるが、この場合は答えが全く違ったものを想定している訳なので別のテンプレートに分けるべきである。この運用中の更新過程は、Burke 達の FAQ-FINDER の WordNet への同意語の追加過程と同等であることに留意されたい。

５　探索過程の流れ

５・１　検索過程の準備

(1) テキスト前処理；ストップワードを削除し、単語、名詞を確認せよ。

(2) システム用語構築：タームリストを構築しグローバル重み係数を与える

(3) 用語(a,q)に対して用語 a が解答部に現れると質問部に用語 q が現れる条件確率と質問部に用語 q が現れると用語 a が解答部に現れる条件確率を求める

(4) 質問と回答を展開し式２を使い FAQ データセット中の質問・解答のタームベクターを構築する

(5) 各 FAQ データセットの各項目にタームベクター l を構築する

(6) 総ての FAQ データセット中の質問・解答の正規化したタームベクターwお構築する

(7) その総ての要素が内部差分タームベクターである内部差分ターム文書行列$D_I^{m\times n_1}$を構築する

(8) その総ての要素が外部差分タームベクターである外部差分ターム文書行列$D_E^{m\times n_2}$を構築する

(9) D_IとD_EをＳＶＤ法（特異値分解法）アルゴリズムを使ってUSVの形に分解し、式３の充度関数P(x｜D_I)とP(x｜D_E)を定義する適当なkの値を求める

(10) 次式を求める

$$P(D_I|x) = \frac{P(x|D_I)P(D_I)}{P(x|D_I)P(D_I) + P(x|D_E)P(D_E)}$$

ここで、P(D_I)には、平均リコール数をデータベース中のドキュメント数で割った値を、P(D_E)は1-P(D_I)の値を与える

５・２　　FAQ 項目の相似度による選別法

DLSI 法によるフィルタリングの過程を以下に説明をする。

a） ユーザの入力クエリを展開する：相似的な意味の範囲値を示すクエリのベクター空間タームベクターの領域に存在する FAQ データセット中の総ての質問・解答項目のタームベクターを、DLSI 法により事後充度関数を用いて見いだす

b） ａで選ばれた各項目についてテンプレートマッチング法(Chen&Tokuda2003)を用いて質問テンプレートと入力クエリ間の相似度を計算し、相似度の最大値を持つ質問項目を求める。この過程で質問テンプレートをユーザの応答に応じて修正する。詳細な手順を次に述べる。

(1) クエリをドキュメントとして扱い、タームとその出現頻度を求め式２によって展開していわゆるタームベクタを求める

(2) 質問ベクターq を正規化する。FAQ データセット中の各タームベクターl を次の３－７の項目に従い処理する。

(3) 差分タームベクター　x=q-l を構築する

(4) 当該ドキュメントの内部充度関数　$P(x \mid D_I)$と外部充度関数 $P(x \mid D_E)$ を計算する

(5) Bayesの事後確率関数　$P(D_I \mid x)$を計算する

(6) 事後確率関数　$P(D_I \mid x)$の値がある閾値(例えば 0.5)以上のドキュメントを選ぶかまたは $P(D_I \mid x)$値の大きい順に決まった数または$P(D_I \mid x)$値の大きい順にＮ個の質問・回答ペアを選ぶ

(7) ステップ６で選択した各 FAQ データセット質問・回答ペアに対して、質問部テンプレートと入力クエリを照合し、最長共通文字列(LCS)を求める

(8) 質問テンプレートと入力クエリの中で最長共通文字列を持つ FAQ データセット中の質問・回答ペアのセットを求める

(9) ユーザにそのテンプレートを示す

(10) ユーザーが承認すれば FAQ データセット中の質問テンプレートを修正する

６　　実験結果

本論文では、Java の９５個の質疑応答を集めた Lucene FAQ コレクション（http://lucene.sourceforge.net/cgi-bin/faq/faqmanager.cgi/）を用いて DLSI 法によるコンテンツ検索・テンプレートによる文字検索による新しい方法の情報検索性能の実験を行った。ターム展開とテンプレートマッチングを用いた。FAQ の回答のどれかに合う５３７個のユーザークエリを英語のネイティブ・スピーカに用意させた。本検索法の各段階毎にどの程度性能が上がるか確認するために次の要領で次の１２段階の実験を行った。

(1) tf-idf 法 (2)ターム展開を行った tf-idf 法　(3)テンプレートマッチングと tf-idf 法

(4)ターム展開とテンプレートマッチングを行った tf-idf 法　(5) LSI 法　(6) ターム展開とＬＳＩ法　(7) テンプレートマッチングと　ＬＳＩ法　(8) ターム展開･テンプレートマッチングと LSI

法　(9) DLSI 法　(10) ターム展開と DLSI 法　(11) テンプレートマッチングと DLSI 法　(12) ターム展開・テンプレートマッチングと DLSI 法。

この実験で我々は、検索項目数は１０、１５、２０個に固定した。図４に精度(precision)＝FAQ 項目の中で正しく検索された項目の数/クエリの全体の数　を示している。各１２個の段階別に示してある。テンプレートマッチングを用いた tf-idf/LSI/DLSI の実験方法について以下に述べておく。まず、tf-idf/LSI/DLSI を使って１０，１５，２０個の項目を選び、それらをテンプレートマッチングによる LCS 値による統語検索を行い、与えられたリストの相似値順のリストを返す。最初の tf-idf/LSI/DLSI によるリストは変わらない訳なので、テンプレートマッチングを行った結果ランキングリストの順序は変わっても検索精度が変わることはない。図４は、12 個の回の中では精度の関して言えばターム展開を伴う DLSI 法が最も高く、次が DLSI 法、ついでターム展開を伴う LSI 法、LSI 法順になっている。図５では、正しい検索結果の平均ランキングの結果が示してある。ここではテンプレートマッチングの威力が大きいことに注意されたい。実用上の問題としては、我々は常に検索結果の上の方におきたいので、このランキングは重要である。

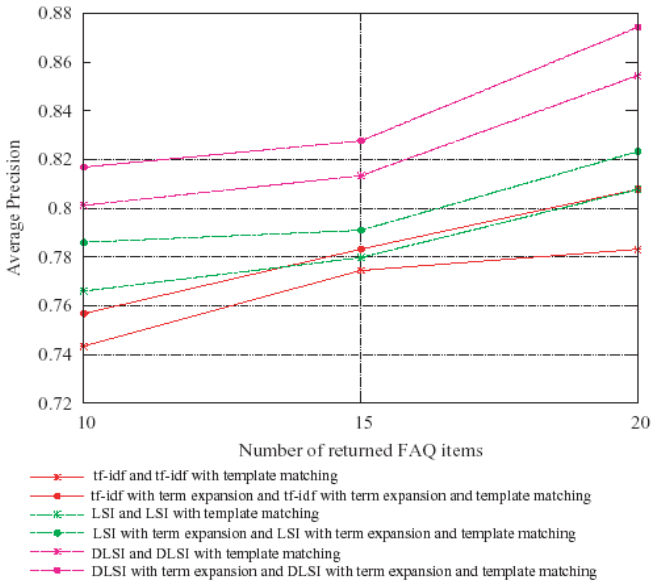

Figure 4: Average precisions of different approaches

図５の結果を見ると正しい検索結果は常にリストの上位に位置していることが分かる。

既に説明したように、テンプレート展開は自然言語の表現数を増やすことを目的にしており、展開されたテンプレートパターンはユーザの自由な表現を数多く組み入れるのに役立っている。本手法でテンプレート展開は非常に重要な役割を果たしていることが解る。この点を確認するために我々は次の実験を行った。ネイティブ・スピーカーが作った 537 個のクエリについて、DLSI 法により FAQ セットから２０組をまず選び、その中からテンプレートマッチングで選んだ最善の１０個を選ぶ。もしそのリスト中に正解のペアがあればそれをテンプレート展開する。537 個のクエリが終われば、又２回目の検索ラウンドを開始する。もしテンプレート展開の効果があるとすれば、検索効率は向上していくはずである。図６に、この何回目かのラウンドに従いどのよ

うに検索効率が動的に向上するかどうかが示されていることに注意されたい。

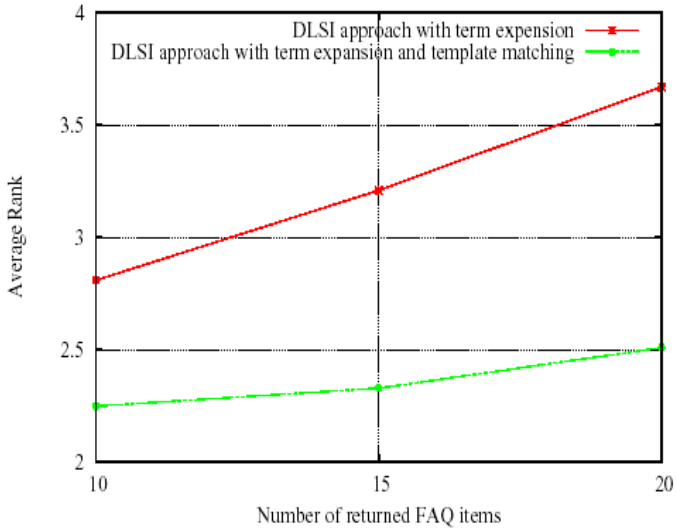

Figure 5: Average ranks of retrieved FAQ items for different approaches

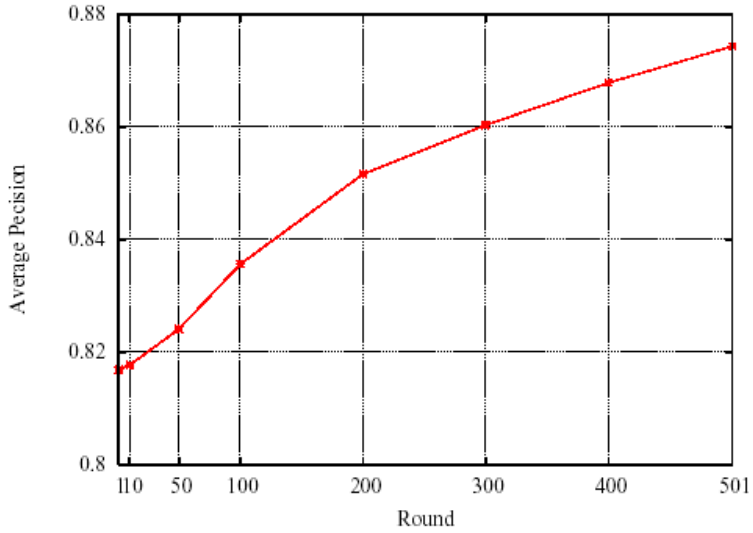

Figure 6: Average ranks of retrieved FAQ items for different approaches

7 討論と結果

本論文では、コンテンツ・統語解析を組み合わせた検索により自然言語による FAQ システムの効率の良い検索方法を開発した。この新しい方法を使うと、意味上の絞り込みと統語上の絞り込みが可能である。質問・回答ペアのタームベクトル間に存在する統語ギャップを除去するために、

ターム展開法によりそのギャップを除去する方法をとった。そうすることにより、DLSI 法で意味的な空間の絞り込みを行い、テンプレートマッチング法により統語検索を行った。この方法は、WORD NET のような編集に手間のかかる同類語辞書を使わなくとも、一番簡単な質問部のテンプレートをユーザが使う自然言語に展開させ、多様な自由度のある質問文を作ることにより発展させていくという手法を取っている。LUCENE という FAQ システムで実験したところ、このテンプレート展開を行うことにより顕著な効果が得られることが判明した。

文献

A.Berger, R.Caruana, D.Cohn, D.Freitag and V.Mittal, “Bridging the Lexical Chasm: Statistical Approaches to Answer Finding.” Proceeding of SIGIR, 192-199, (2000).

R.Burke, K.Hammond, V.Kulyukin, S.Lytinen, N.Tomuro, and S.Schoenberg, “Question answering from frequently-asked question files: Experiences with the faq finder system”, AI Magazine, 18(2): 57-66,(1997).

L.Chen and N.Tokuda “Bug diagnosis by string matching: Application to ILTS for translation”.CALICO Journal, 20(2):227-244, (2003).

L.Chen, N.Tokuda, and A.Nagai “Probabilistic information retrieval method based on differential latent semantic index space.” IEICE Trans. on Information and Systems, E84-D(7):910-914, (2001).

L.Chen, J.Zeng, and N.Tokuda, ” A “Stereo” document representation for textual information retrieval”. JASIST Journal of the American Society for Information Science and Technology, Accepted for Publication, (2005).

L.Chen, N.Tokuda, and A.Nagai, “ A new differential LSI space-based probabilistic document classifier.” Information Proc. Letters, 88(5): 203-212, (2003).

T.R. Gruber. “ Toward principles for the design of ontologies used for knowledge sharing.” Int. Journal of Human-Computer Studies, 43(5-6): 907-928, (1995).

N.Tokuda and L.Chen, ”An online tutoring system for language translation”, IEEE Multimedia, 8(3): 46-55, (2001).

Justin Zobel and Alistair Moffat. “ Exploring the similarity space.” ACM SIGIR FORUM, 32(1): 18-34, (1998).