潜在的トピックモデルに基づく索引語の係り受けを用いた文書検索

柳沢　孝†　三浦　孝夫†

† 法政大学　工学研究科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †takashi.yanagisawa.km@stu.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　近年, 文書の電子化に伴い, 大量の文書データが簡単に入手できるようになっている. このため, 情報検索技術の拡張が数多く提案され, 目的に合致する情報を効率的に発見する試みが論じれらている. しかし利用者の入力する索引語の意図に注目している研究は少ない. 本稿では, 索引語の係り受け関係の潜在的意図を用いて索引語を文脈の一部とした検索手法を提案する. さらに実験でその有用性を検証する.
キーワード　質問, 係り受け, 潜在的ディリクレ分配法

Dependencies of Queries Based on Latent Topic Model

Takashi YANAGISAWA† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †takashi.yanagisawa.km@stu.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　We can capture major quantity of documents easily with computerized documents. Researcher discuss proposed how to many expantion of information retrieval technique and try to efficiently find out information of matching goal. However, few investigation which capture contents of queries directly. We propose how to retrieval dependencies of queries based on latent topic using query as part of context. We show some experimental results to see the effectiveness.
Key words　Query, Dependency, Latent Dirichlet Allocation

1. 前　書　き

近年, 文書の電子化に伴い, 大量の文書データが簡単に入手できるようになっている. すでに人間が全てを読んでそれらを体系化したり, 必要な情報を分類整理することが困難になっている. このため, 情報検索技術の拡張が数多く提案され, 目的に合致する情報を効率的に発見する試みが論じられている.

一般的な検索技術では，質問が出現する文書の情報を転置索引に格納しておくことで，入力される質問と文書の記載事項との合致度により検索するものがある. ここでの質問とは一般的には利用者により複数の単語が与えられるもので, この単語に tf-idf などで重みづけをすることで与えられた単語集合と文書が効率的に合致するものをランク付けする.

一方, 質問に注目した検索技術として, 検索の精度向上を目的とした方法に構文解析データや自然文の質問などから係り受け関係を考慮することで, 高い精度で質問と関係する文書を検索する研究もおこなわれている. [8] [11]

また, トピックモデルのいくつかが情報検索で応用されている. これは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現される. 最近では潜在トピックモデルとして Latent Dirichlet Allocation (LDA) に基づく検索モデルを用いた検索がある. [3]

しかし, 与えられた利用者の意図を表現した索引語集合である質問に対し, その索引語の意図や組み合わせから成る意図を考慮している研究は少ない. 本稿では, 利用者の意図である質問に注目し, 質問内の索引語を含む係り受けを検索に用いて, その潜在的トピックを意図とする検索手法を提案する.

2 章では関連研究，次に 3 章で係り受けの LDA の潜在的トピックモデルに基づく係り受け生成モデルについて述べ, さらに索引語の係り受けの組み合わせから適切な係り受けを選択し, 4 章で実験方法，実験結果，および考察，5 章で結びとする.

2. 関 連 研 究

トピックモデルとは, 一つの文書が複数のトピックの混合として表現されるという仮定である.

一つの文書が一つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化することが可能性がある. その中でも最近用いられているのが LDA である. [1]

確率的潜在意味索引付け (Probablisitic Latent Semantic In-

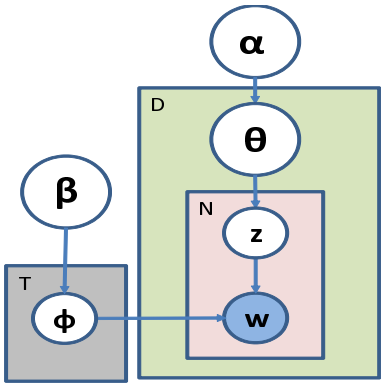

図 1　LDA のグラフィカルモデル

dexing, pLSI) は, LDA と違って, トピックと単語の多項分布をそれぞれにディリクレ事前分布を導入し, 混合比を学習データの文書集合に依存して固定化している.

一方, LDA ではこの混合比は事前分布から動的に生成する点で異なる. LDA は学習データだけに依存しないが, 代わりに特定の確率モデルを仮定するため, 柔軟な観点で文書を扱える可能性がある.

図 1 は LDA のグラフィカルモデルである. 図中の変数は, 図 1 左に, ディリクレ事前分布 $Dir(\beta)$, 図 1 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 1 上にディリクレ事前分布 $Dir(\alpha)$, 図 1 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N は各文書の単語数を表す. LDA の単語生成過程を以下で示す.

まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する. 次に, 文書 d 内の i 番目の単語 w_i において, 抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し, そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する.

Wei らは LDA モデルを用いた検索方法を提案している. [3] 具体的に LDA を以下の式でスムージングとして従来手法に加えて用いている.

$$P(w|D) = \lambda(\frac{N_D}{N_d+\mu}P_{ML}(w|d) + \frac{\mu}{N_d mu}P_{ML}(w|coll)) + (1-\lambda)P_{lda}(w|d)$$

$$P_{lda}(w|D) = \sum_t P(w|t)P(t|D)$$

N_d は文書総数, μ はスムージングパラメタ, $P_{ML}(w|d)$ と $P_{ML}(w|coll)$ はそれぞれ文書 D と文書コレクション $coll$ における単語 w の最尤推定量 (Maximum Likelihood Estimation) により求める. また, $P_{lda}(w|D)$ はギブスサンプリングを用いる.

Wei らの手法では単語の頻度に大きく依存し, LDA はスムージングとしての影響は小さいため, 単語の部分を係り受けとして適用しても, 大量に存在する係り受けの種類を直接適用することは不可能であるため本研究には適さない.

3. 提案手法

本章では, 係り受け関係の意図を表現するモデルと, 索引語を含む係り受けの組み合わせから文書をランク付けする手法について論じる. 係り受けの意図を得るためには係り受けの大量な種類を効率的に LDA に適応する方法が必要である. 本稿では, 係り受けに潜在変数を利用し, 索引語の意図を係り受けから確率的に明確にして適切な文書をランク付けする手法を提案する.

3.1 係り受けと意図

本節では本稿での係り受けと意図の関係について論じる. 本稿での係り受けとは, 係り語 (係り部の文節) と受け語 (受け部の文節) の 2 項関係である. 文節とは日本語で意味の分かる最小単位とされており, 文における任意の 1 つの文節は, 少なくともその文節の後の一つの文節を係り受け関係を持つ特徴がある. [10]

また, 構文解析後, 文節では名詞句など一般的意味を成す語は内容語, 文節内で助詞など文法的な役割を果たす語は機能語としているが, 本稿では名詞集合を質問とする検索を想定しているので助詞以外を内容語として利用する.

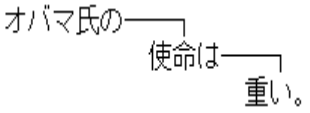

図 2　係り受け解析

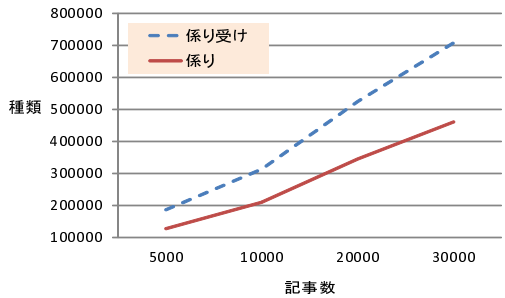

図 3　係り受けの種類

図 2 のように係り語を修飾する受け語が存在し, 両方は最小の意図を表す文節の対となる関係を表している. また, 文法的に先に出てくる係り語は対となる受け語により意味をより明確にする可能性がある. たとえば, 図 2 では, 「使命」を「重い」という語で修飾することでより意味を明確にしている.

しかし, 図 3 のように係り受けすべてを単純に収集するのでは

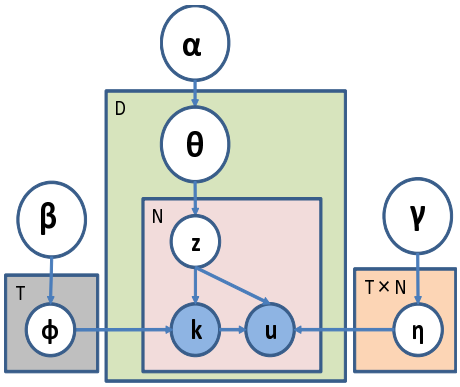

図 4　係り受け LDA のグラフィカルモデル

対の数は大量になる．そこで係り語に対し，それを修飾する受け語を潜在的な状態として扱う．これにより，トピック内の意図と係り語から適切な受け語を確率的に扱うことができる．そこで本稿では独自で係り受け LDA モデルを構築し用いる．

3.2　LDA に基づく係り受け生成モデル

本節では係り受け関係と意図を表現するモデルについて論じる．図 4 に提案モデルを示す．

図 4 では LDA の単語の分布の代わりに新しく係り語と受け語の分布を導入する．図 4 左に，ディリクレ事前分布 $Dir(\beta)$, 図 4 左下は係り語の空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数である．ディリクレ事前分布 $Dir(\gamma)$, 図 4 右下は受け語の空間の多項分布 $Multinomial(\eta_{z_i,k_i})$, T はトピック数, N は係り受けの数とする．図の上部分にディリクレ事前分布 $Dir(\alpha)$, 図 4 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, を表す．以下で係り受けの生成過程を示す．

まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し，トピック t と係り語 k においてディリクレ事前分布 $Dir(\gamma)$ から $\eta_{t,k}$ を抽出する．次に，すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する．最後に，文書 d 内の i 番目の係り受け k_i, u_i において，抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から係り語 k_i とそれに伴う多項分布 $Multinomial(\eta_{z_i,k_i})$ から受け語 u_i を抽出する．

単語生成モデルである LDA との違いは，まず，扱うデータは文書を単語の集合ではなく，係り受けの集合とする．生成されるのは単語ではなく，係り語の多項分布と受け語の多項分布から係り受けとなる．これにより係り受けの潜在的意図を扱う文書モデルが構築できる．また，図 4 のグラフィカルモデルを表す係り受け生成モデルの式を以下に示す．

$$P(k,u|\alpha,\beta,\gamma) = \int\int\int \prod_{z=1}^{T\times N} P(\eta_{z,n}|\gamma) \prod_{z=1}^{T} P(\phi_z|\beta)(P(\theta_d|\alpha)$$

$$(\prod_{n=1}^{N_d}\sum_{z_n}^{T}(P(z_n|\theta)P(k_n|z_n,\phi)P(u_n|z_n,k_n,\eta))))d\theta d\phi d\eta$$

式内の未知パラメタはギブスサンプリングにより推定を行う．図 4 のグラフィカルモデルの式において，分布 $P(\mathbf{z}|\mathbf{k},\mathbf{u})$ に従うサンプルが得られれば θ(文書 d においてトピック j が生成される確率推定), ϕ (トピック j から係り語 $\mathbf{k}$ が生成される確率推定) や η (トピック j と係り語 k から受け語 $\mathbf{u}$ が生成される確率推定) が得られる．これをギブスサンプリングによって求めるため条件確率 $P(z_i = j|\backslash i,\mathbf{k},\mathbf{u})$ を求める．本稿で用いる式を以下に示す．

$$P(z_i = j|\backslash i,\mathbf{k},\mathbf{u}) = \frac{(n^{(u_i)}_{\backslash i,k_i,j}+\gamma)(n^{(k_i)}_{\backslash i,j}+\beta)(n^{(d)}_{\backslash i,j}+\alpha)}{(n^{(\bullet)}_{\backslash i,k_i,j}+\gamma)(n^{(\bullet)}_{\backslash i,j}+\beta)}$$

n^d_j : 文書 d におけるトピック j の出現数.

$n^{(k_i)}_j$: トピック j で係り語 k_iの出現数.

$n^{(u_i)}_{k_i,j}$: トピック j かつ
係り受けが係り語 k_iで受け語が u_iの出現数.

$n^{(\bullet)}_j$: トピック j のコーパス全体での出現数.

$n^{(\bullet)}_{k_i,j}$: トピック j かつ
係り受けが係り語 k_iのコーパス全体での出現数.

上記の式を用いてパラメタの推定を行う．

3.3　係り受けの意図を用いた検索

本節では索引語に基づく係り受けの意図を用いた検索について論じる．検索において，質問は利用者が目的である一つの意図を考え，それを断片的な単語の集合として表現したものと仮定する．この質問内の各索引語にはそれぞれ意図が存在し，それぞれの意図の中で共通となる検索結果が得られればそれが目的の意図を表す結果となる．本稿では，係り受けに置き換え，質問内の語の係り受けにもそれぞれ意図が存在していると仮定し，質問を含む係り受けの共通するトピックを有する文書を検索結果とする．さらに各文書に対し，質問の語を含む係り受け関係の組み合わせについて最尤推定法を用いてランク付けを行う．

4.　実　　験

4.1　実験準備

本稿ではコーパスとして朝日新聞 2009 年 1 月〜12 月の記事数 90880 件うち 1 月 1 日から 10000, 20000 件を用いる．日本語形態素解析システム JUMAN と日本語構文解析システム KNP による解析を行う．また，パラメタは予備実験より決定しておく．

まず，モデルの評価として，モデル内での Topic, 繰り返し回数を決定する．そこで言語モデルの性能評価には一般に，パープレキシティと呼ばれる情報理論に基づく客観的評価手法を用い

る．パープレキシティ値が高いほど語の特定が難しく，言語として複雑である．よって，パープレキシティの値が低いほど言語モデルの性能が高いと評価できる．検証の結果からパープレキシティの値が低いデータを用いる．

トピック内での文書間の関連性を調べるため，各トピックでの文書分布の偏りを調べる．このため，tf-idf 重み付けによる余弦類似度を用いるが，トピック内の文書間の平均余弦類似度と全体での文書間の平均余弦類似度による比較を行う．ただし，ここではデータの疎出現を防ぐため閾値を用いる．

最後に，新聞記事の見出しを用いた検索で手法を変えて比較検証を行う．20000 記事中 10 件を課題とし，正解データは課題の 10 件の見出しを有する文書とする．比較する手法はベースライン手法と提案手法を用いる．ベースライン手法は索引語の tf-idf 用いたベクトル空間モデル [12] [13] の余弦類似度によるランク付けを行う．

ここで用いる提案手法は 20000 記事におけるパープレキシティが最も低いものを用いる．質問は文書の見出しを形態素解析し，その名詞の一部分を用いる．ただし，各質問を抽出した文書の tf-idf 値が高い名詞の上位 2 , 3 で行う．文書のランク付けにおいてどちらが上位に正解データが上がるかを比較する．

4.2 実験結果

	Iteration				
Topic	200	400	600	800	1000
50	6466.88	6184.02	6110.86	6070.93	6054.05
100	5048.64	4873.80	4791.86	4744.52	4721.45
150	4501.88	4333.43	4281.81	4257.28	4238.26
200	4181.79	4073.29	4033.04	4014.49	3993.93

表 1　10000 documents test-set perplexity

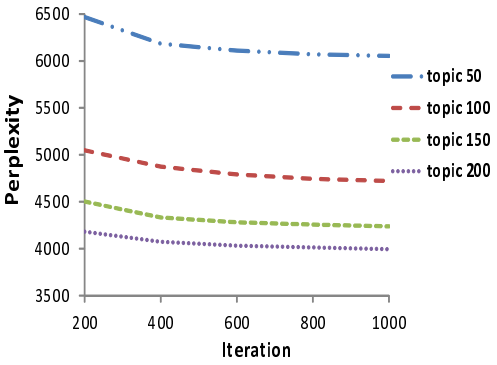

図 5　10000 documents test-set perplexity

繰り返えし回数とトピック数を変え収集した結果を表 1, 2 に示し，さらにグラフでデータを集計したものを図 5, 図 6 に示す．

結果より，パープレキシティは 10000 件も 20000 件も繰り返し回数は 400 回を超えたあたりから徐々に収束し始めている．トピック数に関しては，増やすほどパープレキシティが低下している．以降の実験において，10000 記事においての実験は繰り返し回数 1000 回，トピック数 200 のデータを用いる．同様に，20000 記事においての実験は繰り返し回数 1000 回，トピック数 300 のデータを用いて検証を行う．

	Iteration				
Topic	200	400	600	800	1000
50	7962.55	7639.58	7509.45	7449.90	7418.68
100	6229.89	6026.60	5963.95	5932.45	5905.87
150	5556.85	5373.78	5309.00	5277.59	5260.03
200	5329.33	5176.40	5138.58	5105.77	5071.59
250	5122.78	4989.31	4950.89	4934.12	4920.05
300	5035.64	4931.95	4889.32	4862.85	4856.43

表 2　20000 documents test-set perplexity

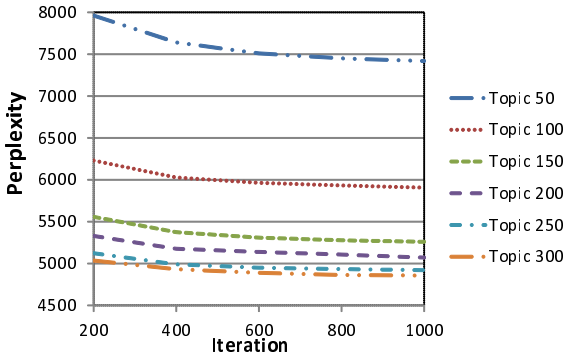

図 6　20000 documents test-set perplexity

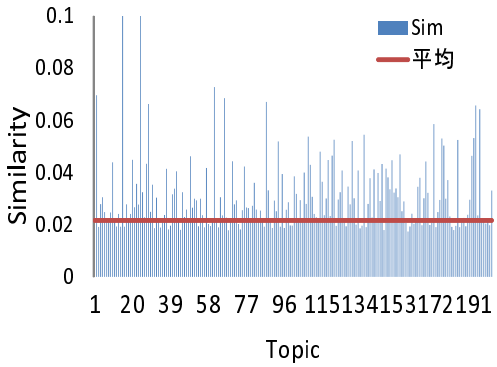

図 7　各トピック内での文書間での平均余弦類似度

topicA	
係り	受け
列車	旅
影響	出た
全線	見合わせ
見合わせ	影響した
売却	検討対象
運転	見合わせ
提案コンペ	実施
証券主要	5社
多額の	損失
反対側	ドア

topicB	
係り	受け
派遣	切り
失職した	農民工
正社員	して
雇用	守る
雇う	いう
労使	合意
短縮する	こと
事業	主に
向き合い	展開すべきだ
派遣期間	上限

topicC	
係り	受け
愛	劇場
根強い	人気
史緒	萌
昭和	におい
男	女
兄	帰る
あり	ふれた
俳優	河原雅彦さん
透き通る	素材
上演	なる

図 8　例

トピック内の文書間類似度は，図 7 に示す．図 7 より，提案手法のトピック内では全体の類似度の平均を 200 トピック中 144 個上回っている．

係り受け LDA における結果の一部として，3 つのトピックの上位 10 個を示す．図 8 よりトピック A は，電車に関する係り受

課題番号	ベクトル空間モデル	提案手法
1	9	6
2	2	6
3	2	4
4	6	3
5	12	3
6	1	8
7	3	1
8	1	15
9	1	1
10	16	4

表 3　単語数 2　でのランク付け

課題番号	ベクトル空間モデル	提案手法
1	7	5
2	2	5
3	2	1
4	6	1
5	12	1
6	1	7
7	3	1
8	1	10
9	1	1
10	16	2

表 4　単語数 3　でのランク付け

け, トピック B においては近年の就職状況に関する係り受け, トピック C では芸能か劇場に関する係り受けが上位に表れている.

文書の検索結果を表 3, 表 4 に示す.　ランク付けの結果の表 3, 4 ではそれぞれ 5 個, 6 個提案手法が tf-idf を用いたベクトル空間モデルの結果を上回っている.

4.3　考　　察

実験に関する考察を行う. 係り受け LDA モデルの評価として, 文書間類似度は, 全トピック内で 72 %が全体の平均余弦類似度を上回っている. 平均余弦類似度を上回る文書を有するトピックの分布が学習により有意に偏っている影響である.

例として, 文書のトピックの分布を左右するのは観測データの係り受けであることから, 平均余弦類似度が 1 位と 200 位のトピックの係り受けのトピックへ所属する確率上位 10 個を図 9 を示す. 図 9 の 1 位の結果では国家政治に関連する係り受け

余弦類似度1位		余弦類似度200位	
係り	受け	係り	受け
突入する	表明	仮眠中	流された
全面対決姿勢	突入する	脱官僚	地域主権
対韓国窓口である	祖国平和統一委員会	斉藤さん	父
南北首脳宣言	07年	兄弟子	3人
動き	ある	同保安部	よる
肯定	否定	第8	き
核保有国	して	かおる	被告
打ち上げる	こと	肖像画	えり
ミサイル	怖いだ	親分	顔
無効	宣言した	相撲	続ける

図 9　平均余弦類似度の 1 位と 200 位の係り受け

が上位に集中している. この係り受けを含む文書同士も類似している結果が得られた. しかし, 図 9 の 200 位の結果では分野が共通しているとは考えにくい係り受けが上位にあり, 結果として各係り受けを含む文書が類似していない. 例のように 72 %の文書のトピック間の分布は係り受け LDA による文書のクラスタリングが有意に機能していると考えられる.

また, 係り語を修飾する受け語の影響によってトピックが変化しているという結果も得られた. 「日本」という係り語について結果を検討する. 2 つのトピック内での「日本」を係り語とする係り受け確率の上位を図 10 に示す. また, 各トピックの上位の係り受けを図 11 に示す.　図 10 では, 各係り受けを見

topicX		topicY	
係り	受け	係り	受け
日本	航空会社	日本	首相
日本	考えられる	日本	訪問団
日本	出発する	日本	支援団
日本	空	日本	とって

図 10 「日本」における係り受け

topicX		topicY	
係り	受け	係り	受け
日本	航空会社	北方	四島
飛行計画	承認	出入国カード	提出
離陸	始めた	戦後	首相
完全に	追い払うの	日本	首相
けが人	無かった	帰属問題	最終的解決

図 11　図 10 の各トピックにおける上位の係り受け

てみると航空関連と国際政治関連の係り受けと見ることができる. 次に図 11 では, 各トピックの上位の係り受けをまとめたものであるが, 各トピック内の係り受けは航空関連と北方領土問題に関連しているものと考えられる. 「日本」におけるトピックが受け語の存在によって変化していることがわかる.

次に, 係り受け LDA を用いた検索では, ベクトル空間モデルと提案手法を比較した結果, 索引語 3 つでは, ベクトル空間モデルの結果を上回る結果となった.

ベクトル空間モデルの質問は課題としている文書の tf-idf 値を用いているため, 高い精度で検索されるはずである. これを考慮すると, 索引語 2 つでの結果も十分に有用性があるといえる.

また, 本手法では質問内の索引語における共通したトピックの係り受けで検索を行っているはずである. そこで, 課題番号 4 における索引語 3 つ用いる実験を例にとって有用性を示す. 見出しの原文は「日露首脳会談：領土問題「政治が決断」　首相,

新アプローチ　【大阪】」, 質問は課題文書内での tf-idf 値に従って「会談」,「首脳」,「露」の 3 つを用いて検索を行った.

各索引語を含む係り受けは図 12 に示す. この索引語の係り受けのトピック内上位 5 個の係り受けを図 13 に示す.

係り	受け
個別会談	行うつもりだ
日露首脳会談領土問題	大阪
日露双方	関心事項

図 12　課題 4 の質問における係り受け

係り	受け
北方	四島
出入国カード	提出
戦後	首相
日本	首相
帰属問題	最終的解決

図 13　課題 4 の係り受けが所属するトピックの上位 5 個の係り受け

図 13 は北方領土問題に関連するトピックであると考えられ, 課題の見出しからすると索引語の係り受け図 12 の意図と合致している可能性が高い.

また, 課題 4 でのベクトル空間モデルの検索において,「会談」,「首脳」,「露」の tf-idf 値はそれぞれ 37.495 ,30.3249, 28.051 であった. 課題 4 のベクトル空間モデルでのランク付けの 1 位の文書の見出しは「日米首脳会談 :「親密さ発信」思惑空振り　昼食会も共同会見もなし」であり,「露」の単語の要素が含まれない文書であった. これは「会談」の tf-idf 値が突出して高いためこの単語の影響により, 他の文書に偏ってしまったと考えられる. 本手法は係り受けとトピックを得ることで, 利用者の質問の意図を考慮した検索により, 語単体の影響はあまり受けなかった. よって, 本手法は利用者の意図である質問から係り受けを用いることで適切な文書検索を行えた.

5. 結　　論

索引語の係り受け関係の潜在的意図を用いて索引語を文脈の一部とした検索手法を提案した. これに伴い, 係り受け LDA を構築し, 文書検索を行った. 係り受け LDA では全トピック中 72 %が文全体の平均類似度を上回り, トピック内の係り受けも潜在的トピックにより偏っている可能性がある結果が得られた. 文書検索では, 重み付けされた従来検索手法を上回り本手法が有用であることを示した.

文　　献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan: "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 993-1022, 2003.

[2] Yexin Wang, Li Zhao, Yan Zhang: "MagicCube:Choosing the Best Snippet for Each Aspect of an Entity.pdf", Proceeding of the 18th ACM conference on Information and knowledge management, 2009.

[3] Xing Wei, W. Bruce Croft : "LDA-Based Document Models for Ad-Hoc Retrieval" Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006.

[4] Takashi Yanagisawa, Takao Miura, Isamu Shioya: Simplifying Sentences by Frequent Parsing Patterns, The 11th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), 2010.

[5] Takashi Yanagisawa, Takao Miura : Sentence Generation for Stream Annoucement, IEEE Intn'l Pacific Rim Conference on Communications, Computers and Signal Processing(PACRIM), 2009.

[6] 鍛治 伸裕, 喜連川 優: " 語彙統語パターンにもとづく制約付き分布クラスタリング", 人工知能学会 知識ベースシステム研究会, 2007.

[7] 江口 浩二, 塩崎 仁博: " 多型トピックモデルを用いた Wikipedia 検索", セマンティックウェブとオントロジー研究会, 2009.

[8] 新里　圭司, 黒橋 禎夫: " 質問の語句の重要度と係り受けを考慮した自然文検索 ", 情報処理学会 IPSJ, 2009.

[9] 黒橋 禎夫: " 結構やるな, KNP", 情報処理 41(11), 1215-1220, 2000-11-15.

[10] 金 明哲: " 統計的テキスト解析 (3)", ESTRELA No170, 2008.

[11] 新見　和彦, 兵藤 安昭, 池田尚志: " 係り受け関係を用いる高精度全文検索 ", 情報処理学会 IPSJ, 1997.

[12] 北　研二, 津田 和彦, 獅々堀 正幹 : " 情報検索アルゴリズム ", 共立出版, 2002.

[13] 徳永　健伸: " 情報検索と言語処理", 東京大学出版会, 1999.