Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

同位語を利用した不在インデックス

新里圭司[†1,*1] 鎌田浩司[†2,*2] 黒橋禎夫[†1]

自然文検索では，文書中に出現する単語，同義語・句，係り受け関係をインデックスに登録し，これらを検索時の文書収集やスコアリングに利用する．しかしながら，クエリ中の語・句の同位語が含まれる文書の適合度を誤りやすいという問題がある．本稿では，文書に「書かれていない」ということを表す不在タームを，国語辞典・ウィキペディアより獲得した同位語を利用して生成し，これを利用することで高速に不適合文書を検出する手法を提案する．

NTCIR-3/4 で構築されたテストセットを用いて提案手法を評価した結果，82.9%の精度で不適合文書を検出できることがわかった．

Non-Existence Index using Coordinate Terms

Keiji Shinzato,[†1] Hiroshi Kamada[†2]
and Sadao Kurohashi[†1]

In natural language search, words, synonyms and dependencies in a document are indexed, and they are exploited for document retrieving and scoring. Natural language search, however, is likely to regard irrelevant documents including coordinate words of terms in a query as relevant ones. To solve the above problem, this paper proposes a *non-existence term* which means that a document does not describe information. For instance, the non-existence term "pigeon → damage" extracted from the document D means that the document D does not describe "damage of pigeon." Non-existence terms are generated by using coordinate words extracted from an ordinary dictionary and Wikipedia, and allow search engines to rapidly detect irrelevant documents.

We evaluated the effectiveness of non-existence terms using the test collection constructed by NTCIR-3/4 competition. Experimental results showed that the proposed method achieved 82.9% in precision for irrelevant document detection.

1. はじめに

"情報爆発" という言葉で形容されるように，現在，ウェブ上には膨大な量の情報が公開されており，それらへの効率的なアクセスを提供する検索エンジンは必要不可欠なライフラインとなっている．現在の検索エンジンは，主にキーワードによる検索であり,「京都大学」や「イチロー」のように企業や人物のトップページを検索する navigational クエリに対しては十分な精度で適合文書を検索できるが,「カラスによる被害を知りたい」のように情報を探索する informational クエリに対しては十分な精度が達成されているとは言い難い．

このような informational クエリに対して精度良く適合文書を検索するための方法として，クエリ中の語・句の修飾関係（係り受け関係）を利用する方法が提案されている[1),2),5),6),11)]．TSUBAKI[*1]は，語・句の修飾関係を利用した自然文による検索が可能であり,「カラス」から「被害」への係り受け関係が出現する文書に対してより高いスコアを与える．

このように係り受け関係を利用することで，より適合度の高い文書を検索することが可能であるが，係り受け関係が出現していない文書に対しては，正しく適合度を判断できないという問題がある．そこで，本論文では文書に**書かれていない内容**に注目することで，問題の解決方法を試みる．具体的には，クエリ中の語・句の修飾関係と検索結果として得られた文書に出現する語・句の修飾関係のズレから，クエリに対して適合しない文書を検出する．情報検索においては検索速度も重要とあるため，本論文ではこのような文書を高速に検出する手法を提案する．

2. 提案手法

2.1 自然文検索における問題とその解決策

2.1.1 解決を試みる問題

自然文検索では，クエリ中の語・句の修飾関係を検索に利用可能である．例えば，TSUBAKI に対してクエリ「カラスによる被害を知りたい」を与えらと図 1 に示す文書が得られる．図 1 における文書 A〜F はいずれもクエリ中の内容語「カラス」「被害」を含んでいる．

†1 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University.
†2 京都大学工学部電気電子工学科, School of Electrical and Electronic Engineering, Kyoto University
∗1 現在，楽天技術研究所.
∗2 現在，NTT コミュニケーションズ.
⋆1 http://tsubaki.ixnlp.nii.ac.jp

Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

○	文書 A	...**カラス**による**被害** が急増しており，住民を悩ませている．例えば，ゴミの散乱...
○	文書 B	... フン害など**カラス**に代表される 鳥の**被害** が区役所に寄せられている．...
○	文書 C	... **カラス**の鳴き声で睡眠を妨げられてしまう「騒音**被害**」が一番の...
×	文書 D	... 近くの工事現場による 騒音**被害** のおかげ？で，**カラス**の数が激減しました．...
×	文書 E	...ハトによる**被害** を減らすには**カラス**などの天敵を模してぶら下げよう...
○	文書 F	...ハトによる**被害** よりも，まだまだ**カラス**の方が問題でしょう．例えば鳴き声...

図 1 クエリ「カラスによる被害」の結果文書 (○×は文書がクエリに対して適合か不適合かを表す)

文書 A は，クエリと同様に「カラス」から「被害」への係り受け関係（下線部）を含んでおりクエリに対して適合している（以下，語 A から語 B への係り受け関係を「A → B」で表す）．一方，残りの文書には「カラス→被害」は含まれておらず，そのかわりに「鳥→被害」「騒音→被害」「ハト→被害」が含まれている（下線部）．

クエリに含まれる係り受け関係の出現を調べることで，文書 A の適合度が高いと判断し，上位にランキングすることが可能である．しかし，残りの文書（文書 B〜F）についてはクエリ中の係り受け関係が文書内に出現していないため，係り受け関係の出現を手掛かりに文書の適合度を判断できない．この場合，文書の適合度はクエリ中の単語やその同義語の出現頻度，文書長，被リンク数などを手掛かりに判断される．そのため，文書 E のように「カラスの被害」について明らかに書かれていない不適合文書がランキングの上位にくる可能性があり，これが検索精度の低下を招いている原因の 1 つと考えられる．

2.1.2 不適合文書を検出するためのアイデア

図 1 の文書 B〜E を見ると「被害」を修飾している語に注目することで，文書の適合度が判定できそうである．文書 B は「カラス」の上位語である「鳥」の被害について書かれたものであるが，その中でカラスの被害について書かれている可能性が考えられるため，一概に適合度が低い文書と見なすことはできない．文書 C, D は「騒音」の被害について書かれているが,「カラス」や「工事」など様々な騒音の被害が考えられるため,「騒音→被害」の係り受け関係から適合度の判断はできない．しかしながら，文書 E は「カラス」と同位関係にある「ハト」の被害について書かれたものであり，これは適合度が低いと見なせる．同位関係とは，**共通の上位語を持つ表現同士の関係** であり,「カラス」「ハト」は共通の上位語「鳥」を持つため同位関係にある．先に述べた，文書 C，D の「騒音」は「カラス」と共通の上位語を持たないため同位語ではない．同位語同士は類似した文脈に出現しやすいという特徴がある.「カラスによる被害」で検索された文書中に「カラス→被害」ではなく「ハト→被害」が含まれている場合，この文書がカラスによる被害について書かれている可能性は非常に低く不適合文書と見なせる．その一方で,「騒音」と「カラス」のような同位関係にない語同士の場合は,「騒音→被害」がカラスの被害の意味を表している可能性があるため，不適合文書とは断定できない．

以上より，クエリ中の係り受け関係（カラス→被害）の係り元（もしくは係り先）が同位語に置き換わった関係（ハト→被害）を含んでいる文書は適合度が低いと仮定し，このような係り受け関係を含んでいる文書を不適合文書と見なす．もちろん文書 F のように「ハト→被害」を含む適合文書も考えられるが，このような例は少数であると考えられるため，不適合と判断してしまって問題ない．

2.1.3 検出を高速化するためのアイデア

「検索」には高速な応答が求められるため，膨大な量ある検索結果から素早く同位語に置き換わった係り受け関係を含む文書を検出する必要がある．このような係り受け関係を含む文書を検出する単純な方法として，例えば検索結果中の文書を 1 件ずつ解析して調べる方法が考えられる．しかしながら，一般に検索結果として大量の文書が得られるため，このように 1 件ずつ解析する方法では高速な応答が得られない．

次に，クエリ中の係り受け関係の係り元（または，係り先）を同位語で置き換えた係り受け関係を自動生成し，それらを検索に利用することで不適合文書を検出する方法が考えられる．しかしながら，一つの語に対して同位語は数十個程度考えられるため，検索に利用する係り受け関係の数も同様に増加する．そのため，この方法では通常検索の数十倍の検索時間を要することになり，高速な応答が得られない．

そこで本研究では，**書かれていない情報**を事前にインデックスに登録することで検索と同時に文書が同位語に置き換わった係り受け関係を含むかどうかを検出する手法を提案する．先程の例であれば，文書 E がカラスの被害について書かれていないということを表すターム（**不在ターム**）をインデックスに登録することで検索と同時に不適合文書を高速に検出できるようになる．

2.2 拡張同位語の獲得

同位語とは，**共通の上位語をもつ表現**のことであり，例えば，図 2 の「みかん」に対する「桃」「バナナ」「リンゴ」がそれにあたる．本手法では，従来の同位語の定義を拡大し，同

Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

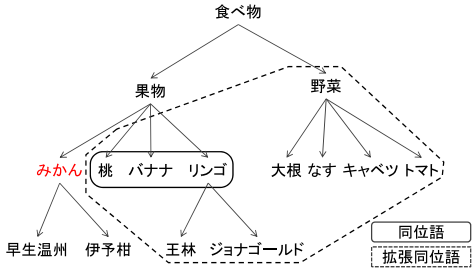

図 2 同位語，拡張同位語の例

じシソーラスに含まれる表現の中から上位語，下位語を除いた表現を同位語と見なし**拡張同位語**と呼ぶ．図 2 の例でいえば,「みかん」に対して従来の同位語に加え,「大根」「なす」「キャベツ「トマト」「王林」「ジョナゴールド」を拡張同位語と見なす．

拡張同位語を獲得するためには，多種多様な表現を含むシソーラスが必要である．そこで，国語辞典および日本語版ウィキペディアを利用してシソーラスの自動構築を試みる．まず，国語辞典，ウィキペディアより表現間の上位下位関係を抽出し，それらを組み合わせてシソーラスを構築する．具体的には，見出しとその定義文 (第 1 文) を抽出し，定義文に含まれる上位語を獲得することで上位下位関係を獲得する．上位語の獲得は，Shibata らが国語辞典から上位下位関係を抽出する際に利用したパターンを用いる[10]．例えば，以下の定義文[⋆1]からは “植物> 野菜”，“野菜>チンゲンサイ”，“キンセンカ>園芸植物” が抽出される（抽出される上位語には下線を引いている）．

野菜 大根・ねぎ・トマト・キャベツなど，食用にするために畑で育てる 植物.

チンゲンサイ チンゲンサイ（青梗菜，学名：Brassica rapa var. chinensis）は，アブラナ科の野菜．中国野菜の中でも身近な 野菜 のひとつとなっている．

キンセンカ キンセンカ（金花，学名：Ｃａｌｅｎｄｕｌａ ｏｆｆｉｃｉｎａｌｉｓ）は，キク科の 園芸植物.

獲得された上位語が複合名詞の場合は，その主辞をさらに上位語として獲得する．例えば「キンセンカ」の定義文からは，“植物>園芸植物>キンセンカ” という上位下位関係が獲得される．

次に抽出した上位下位関係を利用してシソーラスを構築する．例えば，“植物>野菜”，“野菜>チンゲンサイ” という関係が抽出された場合，それらを組み合わせ，“植物>野菜>チンゲンサイ” とする．この処理を抽出された全ての上位下位関係に対して行うことでシソーラスを構築する．その際，上位語が多義語の場合は上記の処理を行わない．多義語かどうかの判定には日本語形態素解析器 JUMAN[⋆2]の辞書を利用する．JUMAN の辞書には，多義である見出し語に対して，人手でラベルが付与されている．また，上位語が,「こと」や「もの」のように抽象的すぎると，あらゆる表現の間に同位関係が成立してしまう．そのため，構築したシソーラスの上位語の抽象度をウェブ 1 億件から取得した文書頻度で近似し，文書頻度が 100 万以上の表現は抽象的すぎると見なしシソーラスから削除する．

このようにして構築したシソーラスより拡張同位語を獲得する．

2.2.1 拡張同位語の整理

国語辞典およびウィキペディアの定義文を利用して上位語を獲得するということは，各表現に対して 1 つの側面からみたときの上位語を獲得しているに過ぎない．実際は表現に対して複数の上位語を考えることができる．例えば「梅」の上位語として定義文から「木」が獲得されているが，他にも「果樹」や「園芸植物」などが考えられる．しかしながら，単に辞書等からシソーラスを構築しただけでは，これらの表現は「梅」の上位語ではなく，拡張同位語として扱われてしまう可能性がある．そこで，拡張同位関係にある語と語に対して，上位下位関係が成り立っているかどうかチェックし，成り立つ場合は拡張同義語と見なさないようにする．

上位下位関係のチェックには，以下の語彙統語パターンを利用した．

- $\{X(\text{や} \mid \text{と} \mid \cdot)\} * X \{ \text{などの} \mid \text{等の} \} Y$
- $\{,X\} * \{ \text{などの} \mid \text{等の} \} Y$

具体的には，拡張同位関係にある語 X と Y を上記のパターンにあてはめ，ウェブ文書 1 億件中に一度でも出現した場合，上位下位関係にあると見なし削除した．

2.2.2 拡張同位語データベース

例解岩波国語辞典および日本語版ウィキペディアに対して前節で述べた手法を適用し，約 40 万表現に対して拡張同位語が登録されている拡張同位語データベース（約 23GB）を構築した．データベースには，1 表現に対して平均約 4,400 個の拡張同位語が登録されている．

⋆1 例解岩波国語辞典，2010 年 3 月時点でのウィキペディアより抜粋.

⋆2 http://nlp.kuee.kyoto-u.ac.jp/nlp-resorces/juman.html

Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

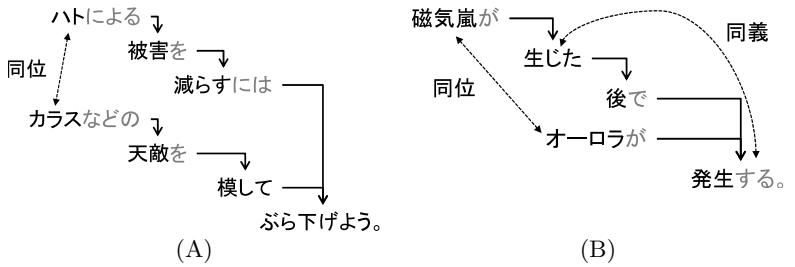

図 3 構文解析結果

2.3 不在タームの生成

2.2.2 節で構築した拡張同位語データベースを利用し，インデキシングする際に不在タームを生成する．具体的には，係り受けターム（A → B）をインデックスに追加する際に，その係り元 A（または，係り先 B）の拡張同位語 C が近傍に出現している場合，新たに不在ターム C → B（または A → C）を生成する．本手法では，係り元または係り先の語の近傍 100 語以内にどちらかの拡張同位語が出現している場合，不在タームを生成する．

不在タームをインデックスに登録する際，不在タームと同じ係り受けターム，もしくは同義と見なせる係り受けタームが同一文書に存在しているかどうか確認し，存在している場合は登録しない．係り受けタームの同義性は，国語辞典より構築した同義語・句データベース[10] を用いて判定する．また，複合名詞については，その主辞のみを用いた係り受け関係を不在タームとして登録する．

例として，以下の文から不在タームを生成することを考える．

(a) ハトによる被害を減らすにはカラスなどの天敵を模してぶら下げよう (図 1 の結果文書 E)

(b) 磁気嵐が生じた後でオーロラが発生する．

文 (a), (b) の構文解析結果を図 3 に示す．図 3 (A) の場合,「ハト」が「被害」を修飾することから係り受けターム「ハト→被害」が得られる.「ハト」の近傍には拡張同位語「カラス」が出現しているため,「カラス→被害」が不在タームとして生成される．その一方で,「カラス」に注目すると，係り受けターム「カラス→天敵」から不在ターム「ハト→天敵」が生成される．

図 3 (B) からは，係り受けターム「磁気嵐→生じる」「オーロラ→発生」が生成され，さらに「磁気嵐」と「オーロラ」が同位関係にあることから，不在ターム「磁気嵐→発生」と「オーロラ→生じる」が考えられる．しかしながら,「発生」と「生じる」が同義であるため,「磁気嵐→生じる」と「磁気嵐→発生」,「オーロラ→発生」と「オーロラ→生じる」がそれぞれ同義と判断され，実際には不在タームは生成されない．

2.4 文書のスコアリング

クエリと文書の関連度の計算には Okapi BM25[7] を利用する．Okapi BM25 は，クエリ中の単語と文書の関連度を計算するために用いられるが，本研究では単語だけでなく，係り受け関係，不在タームを考慮して関連度を計算する．T_{q_w} をクエリ q から抽出された単語の集合，T_{q_d} を係り受けの集合としたとき，文書 d とクエリ q の関連度 $Rel(q,d)$ を以下の式で求める．

$$Rel(q,d) = \alpha \sum_{t \in T_{q_w}} BM(t,d) + \beta \sum_{t \in T_{q_d}} BM(t,d) - \gamma \sum_{t \in T_{q_d}} BM(t,d)$$

ここで α，β，γ はスコアリングに単語，係り受け関係，不在タームをどの程度考慮するかを調節するパラメータである．実験では $\alpha = 0.8$，$\beta = 0.2$，$\gamma = 0.2$ とした．$BM(t,d)$ は以下の式で定義される．

$$BM(t,d) = \frac{(k_1+1)F_{dt}}{K+F_{dt}} \times \frac{(k_3+1)F_{qt}}{k_3+F_{qt}}$$
$$w = \log \frac{N-n+0.5}{n+0.5}, K = k_1((1-b)+b\frac{l_d}{l_{ave}})$$

F_{dt} は文書 d 中での t の出現頻度，F_{qt} は q 中での t の出現頻度，N は検索対象となっている文書数，n は q の文書頻度，l_d は d の文書長（単語数），l_{ave} は平均文書長である．また，k_1,k_3,b は Okapi のパラメータであり，$k_1 = 1$,$k_3 = 0$,$b = 0.6$ としている．

3. 評 価 実 験

3.1 評価セット

NTCIR-3, 4[3),4)] で構築されたテストセットを利用して評価実験を行った．NTCIR-3,4 では同じ文書セット（jp ドメインから収集された 11,038,720 文書）に対して検索課題が用意されており，課題数は NTCIR-3 が 47，NTCIR-4 が 80 である．検索課題ごとに,「高適合」「適合」「部分的適合」「不適合」「未判定」の 5 段階評価が文書に付与されており，実

Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

表 1 不整合文書の検出精度

	不適合文書の検出精度 [%]	
データセット	ベースライン	提案手法
NTCIR-3	**86.15(13530/15704)**	77.27(34/44)
NTCIR-4	80.45(18134/22549)	**90.63(29/32)**
NTCIR-3/4	82.80(31695/38275)	**82.89(63/76)**

表 2 各評価セットのスコア（−：不在ターム無，＋：不在ターム有）

データセット	MRR-10	P@10	R-prec	MAP	DCG
NTCIR-3(−)	**0.4462**	**0.2404**	0.1628	**0.1188**	**12.4044**
NTCIR-3(+)	**0.4462**	0.2383	**0.1630**	**0.1188**	12.3809
NTCIR-4(−)	**0.4911**	**0.2712**	**0.1770**	**0.1249**	14.9181
NTCIR-4(+)	**0.4911**	**0.2712**	**0.1770**	**0.1249**	**14.9184**
NTCIR-3/4(−)	**0.474**	**0.260**	**0.172**	**0.123**	**13.988**
NTCIR-3/4(+)	**0.474**	0.259	**0.172**	**0.123**	13.979

検索クエリ 「イスラエル とパレスチナ間の 紛争 について何故起こったのか知りたい」
文書 ・イスラエル がレバノン南部からの撤退を早めるなか，国連レバノン暫定軍（UNIFIL）は，イスラエル防衛軍とその同盟軍が撤退した地域の状況について,「予想に反して，静穏である」と報告した．
・プレンダーガスト政治問題担当事務次長が，エリトリア・エチオピア紛争 に関する安保理非公式協議において，ブリーフィングを行った．その後，安保理議長は報道声明を発表し，安保理がこれまでに採択した 2 つの決議に従って，同紛争の両当事者が即時停戦に合意するよう促した．

図 4 不在タームを含む不適合文書の例 1

検索クエリ 「京都の寺や神社 について，歴史的背景，地域での存在など，一歩踏み込んだ情報を知りたい」
文書 氏郷二十六歳の天正十年本能 寺 の変で織田信長が憤死した時，安土城にいた信長の妻妾一族をこの城へ迎え入れ明智光秀の軍を迎え撃とうとしたことは有名である．... 黒川を若松と改め九十二万七千石の大名となったが，文禄四年四十歳の若さで 京都邸 に於て病没した．

図 5 不在タームを含む不適合文書の例 2

験では「高適合」「適合」「部分的適合」を正解,「不適合」「未判定」を不正解とみなした．

実験には<DESC>タグで囲まれた自然文を検索クエリとして利用した．各クエリについて検索結果の上位 1000 件を評価対象とした．

3.2 不整合文書の検出精度

127 クエリについて不在タームを考慮した検索を行い，上位 1000 件に含まれる不在タームを含む不適合文書の割合（不適合文書の検出精度）を調査した．その結果を表 1 に示す．表中の「ベースライン」は，不在タームを考慮しない場合に検索結果上位 1000 件に含まれる不適合文書の割合である．表より，NTCIR-3 についてはベースラインが，NTCIR-4 では提案手法の方が不適合文書検出の精度が高いことがわかる．また，NTCIR-3，4 の両方を考慮した場合は，提案手法が 82.89%，ベースラインが 82.80%であり，提案手法がベースラインを僅かに上回った．以上の結果より，検索結果から無作為に文書を選び不適合文書と見なすよりも，不在タームを手掛かりに不適合文書を選択する方が，より高い精度で不適合文書を検出できることがわかる．

3.3 情報検索システムとしての評価

続いて，情報検索でよく利用される評価尺度を用いて，不在タームを考慮する場合と考慮しない場合について評価を行った。評価尺度としては，MRR@10，P@10，R-prec，MAP，DCG を利用した．ここで，MRR@10 は検索結果の上位 10 件を対象に求めた MRR 値を，P@10 は上位 10 件における適合文書の割合（精度）である．評価結果を表 2 に示す．表より，不在タームを考慮しても各評価尺度に大きな変化が見られないことがわかる．本手法を用いて各評価尺度のスコアが向上するためには，不適合文書のスコアを下げることで，さらに下位にあった適合文書のランクが上昇しなければならない．そのため，現在よりも不適合文書を高精度で検出し，さらに不在タームが含まれる文書のスコアをより大きく下げる必要があると考えられる．

4. 考　　察

4.1 不在タームを含む不適合文書

図 4，図 5 に不在タームを含む不適合文書の例を示す．図 4 は,「イスラエル」と「レバノン」の問題について記述されているが,「イスラエル」と「パレスチナ」の紛争についての記述はない．その代わりに「エチオピア」との紛争について記述されており，クエリに対して不適合な文書である.「イスラエル」と「エチオピア」は「国家」を上位語に持つため同位関係にあり,「エチオピア→紛争」から不在ターム「イスラエル→紛争」を生成することで，提案手法はこの文書を不適合文書として検出できた．

図 5 は，滋賀県にある中野城に関する歴史を紹介した文書であるが，クエリ中の内容語「京都」「寺」が文書内に出現しているため検索結果として得られる．文書内には係り受け関係「京都→邸」が出現しており（下線部）,「邸」と「寺」(二重下線部) の同位関係を介して「京都→寺」が不在タームとして生成され，不適合文書として認識することができた．

Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

検索クエリ　「観測のために オーロラ の 発生 する条件が知りたい.」
文書　太陽フレアと呼ばれる太陽表面の爆発が起きると，その 2〜3 日後に地球に 磁気嵐 が 発生 し，それに伴って オーロラ活動もさかんになります.

図 6　不在タームを含む適合文書の例 1

検索クエリ　「キリストの復活を 祝う「イースター」の祭りについて書かれている文書を探したい」
文書　イースターは，キリストが復活した事を祝う 宗教的な要素が色濃いですが，宗教が広まる以前から...キリストの誕生を祝うクリスマス は一見おめでたいようですが，...

図 7　不在タームを含む適合文書の例 2

4.2　不在タームを含む適合文書

図 6 に示した文書は，拡張同位関係にある表現「オーロラ」と「磁気嵐」が出現しており，下線部の「磁気嵐→発生」に対して，不在ターム「オーロラ→発生」が生成される．文書には，係り受け関係「オーロラ→発生」は出現していないが，それと同様の内容が記述されており（二重下線部），この文書は適合文書である．しかしながら,「発生」と「活動がさかんになる」が同義であることが把握できていないため，結果として，不在ターム「オーロラ→発生」が生成されてしまった．この問題を解決するためには，係り受けタームの同義性判定処理をより高度化する必要があると考えられる．

図 7 に示した文書は,「イースター」だけでなく「クリスマス」についても記述されているが適合文書である．この文書を誤って不適合と判断してしまう理由は，クエリから得られる係り受け関係「祝う→イースター」と，文書から得られる「イースター→祝う」（二重下線部）を同義として認識できていないためである．結果として,「イースター」と「クリスマス」の同義関係，および係り受け関係「祝う→クリスマス」から不在ターム「イースター→祝う」が生成されてしまい，この文書のスコアを下げてしまった．この場合，係り受けタームの正規化などが必要であると考えられる．

他には，シソーラスから得られる拡張同位語が適切でないために適合文書のスコアを下げてしまっている例がみられた．例えば，拡張同位語データベースでは,「加速器」と「機関」が同位関係にあると見なされており，これは「機関」を多義語として扱えていないことが原因であった．

5. 関 連 研 究

情報検索の精度を向上させるため，クエリを構文解析することで求まる語・句の係り受け関係を利用する手法が提案されている．Katz ら[1] や Shinzato ら[11] は，以下のように，同じ内容語から構成される意味の異なる表現を区別するため，語・句の係り受け関係をタームとしてインデックスに登録している．

文書 1　ドイツ車を日本へ輸入する
文書 2　日本車をドイツへ輸入する

例えば，クエリ「日本へ輸入される車」で検索した場合を考える．クエリ中の単語「日本」「輸入」「車」は，文書 1・文書 2 ともに含んでいるが，クエリ中の係り受け「日本→輸入」が出現している文書 1 を文書 2 よりも高く順位付けすることが可能となる．また，Miyao ら[6]，Shen ら[2]，Bilotti ら[5] は，構文解析よりも詳細な述語項構造解析や Semantic role labeling の結果を情報検索に用いることで検索精度の向上を図っている．これらの研究は，クエリや文書に含まれる語・句の修飾関係をタームとして利用することで，検索精度の向上を図っている．しかし，利用する情報はクエリや文書に存在する内容だけであり，本研究のように文書にない情報を扱い，事前にインデックスとして登録するようなことは行っていない．

一方で，クエリや文書中に書かれていない内容を獲得し利用する手法としては，疑似適合性フィードバックという手法が提案されている．疑似適合性フィードバックでは，クエリの拡張に用いた語・句を含む文書ほど適合度が高いという仮定を設けるものが一般的であるが，反対に拡張に用いた語・句を含む文書ほど適合度が低いという仮定を置く手法 (Negative relevance feedback) も存在する．Yan ら[8),9)] は，初期検索結果のランキングが最下位の結果を負例とし，その結果中に含まれる語・句を含むものは適合度が低いと仮定して再検索する手法を提案している．この手法では，検索を複数回行う必要があるため，時間コストが増大するという問題がある．また，検索結果の最下位の文書が必ずしも不適合文書とは限らない．本手法は，初期検索を必要とせず，また，不適合文書に出現しやすい語・句の修飾関係を不在タームとして認識しており Yan らの手法とは異なる．

Vol.2011-NL-201 No.15
Vol.2011-SLP-86 No.15
2011/5/17

6. おわりに

本論文では，自然文検索において同位語と係り受け関係を利用することで，適合度の低い文書を高速に検出する手法を提案した．具体的には，クエリ中の語・句の修飾関係と検索結果として得られた文書に出現する語・句の修飾関係のズレから，クエリに対して不適合である文書を検出する．このような文書を高速に検出するために，文書から抽出した係り受けタームと，自動構築したシソーラスから得られる同位関係を利用して，**文書に書かれていない内容**を事前にインデックス（不在インデックス）に登録する．

NTCIR-3，4 で構築されたテストセットを利用して提案手法の評価を行った結果，不在タームを利用した不適合文書の検出精度は 82.9 %であった．一方で，情報検索システムの評価に用いられる評価尺度にはあまり変化が現れなかった．情報検索で利用される評価尺度において，より大きな値を得るためには，不在タームの生成方法，ランキング時の不在タームの利用方法を，さらに検討する必要がある．

今後の課題としては，提案手法では，不在タームを生成する際に同義の係り受けタームが文書から抽出されていないかどうかチェックしているが，この同義性のチェックをより高度化することが考えられる．また，国語辞典，ウィキペディアをもとに自動構築したシソーラスより同位語は抽出されるため，このシソーラスの整理・拡充も今後の課題である．

参考文献

1) B.Katz and J.Lin: Selectively using relations to improve precision in question answering, *Proceedings of the EACL-2003 Workshop on Natural Language Proceedings for Question Answering*, pp.43–50 (2003).
2) D.Shen and M.Lapata: Using Semantic Roles to Improve Question Answering, *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.12–21 (2007).
3) Eguchi, K., Oyama, K., Aizawa, A. and Ishikawa, H.: Overview of web task at the fourth ntcir workshop, *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization* (2004).
4) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: The web retrieval task and its evaluation in the third ntcir workshop, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2003).
5) M.Bilotti, P.Ogilvie, J.E.: Structured Retrieval for Question Answering, *Proceeding of the SIGIR-2007*, pp.351–358 (2007).
6) Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. and Tsujii, J.: Semantic retrieval for the accurate identification of relational concepts in massive textbases, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp.1017–1024 (2006).
7) Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A. and Lau, M.: Okapi at TREC, *Text REtrieval Conference*, pp.21–30 (1992).
8) Rong Yan, Alexander G.Hauptmann, R. J.: Multimedia Search with Pseudo-relevance Feedback, *Proceeding of the CIVR2003*, pp.238–247 (2003).
9) RongYan, AlexanderG.Hauptmann, R.J.: Negative Pseudo-Relevance Feedback in Content-based Video Retrieval, *Proceeding of the ACM-2003*, pp.343–346 (2003).
10) Shibata, T., Odani, M., Harashima, J., Oonishi, T. and Kurohashi, S.: SYNGRAPH: A Flexible Matching Method based on Synonymous Expression Extraction from an Ordinary Dictionary and a Web Corpus, *Proc. of IJCNLP2008*, pp. 787–792 (2008).
11) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, *Proc. of IJCNLP2008*, pp.189–196 (2008).