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

2005－FI－80 (19)
2005－NL－169 (19)
2005／9／30

土木関連用語辞典の見出し語の分析と検索システムにおける活用に関する考察

相澤 彰子（国立情報学研究所／総合研究大学院大学）、野末 道子（鉄道総合研究所）、
今 尚之（北海道教育大）、坂本 真至（土木学会図書館）、中渡瀬 秀一（総合研究大学院大学）

連絡先：相澤彰子〒101-8430 東京都千代田区一ツ橋 2-1-2
e-mail: aizawa@nii.ac.jp

専門的ポータルサイトにおける検索ナビゲーション機能実現のための索引語抽出について検討する。特に、(i) 分野に特化した用法に対応できること、(ii) 言語処理の専門家でない一般のシステム管理者が容易に辞書管理が行えること、の2点を重視して、既存の専門用語の見出し語集合から単語nグラムどうしの連接コストを定め、これに基づき複合語内解析を行う手法を検討する。また、土木関連分野の専門的な用語辞典の見出し語集合を用いた予備的な結果を報告する。

Analysis of an Entry Term Set of a Civil Engineering Dictionary and Its Application to Information Retrieval Systems

Akiko AIZAWA (National Institute of Informatics / The Graduate University for Advanced Studies),
Michiko NOZUE (Railway Technical Research Institute),
Kon NAOYUKI (Hokkaido University of Education),
Shinji SAKAMOTO (Library of Japan Society of Civil Engineers),
Hidekazu NAKAWATASE (The Graduate University for Advanced Studies)

This paper focuses on the problem of index term extraction for retrieval navigation with domain-specific portals. Targeting to the following two points, (i) domain-specific usage of terms can be easily taken into account, and (ii) administrators of systems can maintain the necessary linguistic resources without specific knowledge on NLP, we propose a new method for analyzing compounds and multi-words based on the cost table obtained from the collection of entry terms of existing term dictionaries. Results of some preliminary experiments are also reported where the proposed method is applied to terms in civil engineering field.

1 はじめに

本稿では、土木関連用語に焦点をあてて、専門的ポータルサイトにおけるテキストからの索引語自動抽出の問題について検討する。学術分野における専門的な用語の大半は、複数の形態素から構成される「複合語」であるが、文献 [1] でも指摘されているように、何を複合語切り出しの単位とするかという問題は、アプリケーション依存の側面を持つため統一的な扱いが困難で、汎用的な分かち書き処理による対応がむずかしい。

具体例をあげると、「高レベル放射性廃棄物」という語は用語として1つのまとまりをなすが、形態素解析[1]の結果では、「高」＋「レベル」＋「放射」＋「性」＋「廃棄」＋「物」という6つの形態素の連なりとなる。このような場合に、(i) 最長単位である「高レベル放射性廃棄物」だけを索引語とする、(ii)「高」、「レベル」、「放射」、「性」、「廃棄」、「物」の単語を索引語とする、(iii) 両者を併用する、(iv)「高レベル」「廃棄物」といった中間的な構成語も索引語とする、の選択肢があり、さらに最後のケースでは、「レベル放射」などの不適格な語をどうするのかという問題が生じる。

[1] http://chasen.naist.jp/hiki/ChaSen/

我々はこれまでに、土木学会附属図書館[2]の目録・書誌検索サービスの機能向上について検討し、連想検索エンジンGETA[3]を使った検索語ナビゲーションの有効性について、利用者アンケートを含めた評価を行ってきた[2]。GETAの連想検索機能を用いると、検索対象文書から抽出した索引語を、検索語候補として画面上に表示して利用者を支援することが可能であるが、この際に「レベル放射」のように不適切な語が提示されると、システムへの使い勝手を損なうことになる。したがって、このような応用において、テキストから適切な単位の複合語を切り出すことはとりわけ重要である。

ここで注意が必要なのは、専門的ポータルサイトを構築しようとする管理者は、必ずしも言語処理の専門家ではないことである。たとえば、形態素解析ツール辞書への用語登録によって、望みの解析結果が得られる可能性もあるが、登録語に対する重みの調整は容易ではない。また、「廃棄物」と「放射性廃棄物」の両方を同時に切り出すといった処理は、単純に既存のツールを適用するだけでは実現できない。現状では、複合語への対応はサイト毎のヒューリスティックに頼らざるを得ないのが実情であり、管理者は、自前の辞書や処理ソフト作成の負担を強いられることになる。

以上の背景を踏まえて本稿では、既存の用語辞典等による専門語の見出し語集合を分析し、その結果を利用して複合語解析および索引語抽出を行う方法を検討する。以下、**2.** で、日本語の複合語抽出問題に関する関連研究を概観し、本稿におけるタスク設定を述べる。**3.** で、用語集の見出し語解析結果を利用した索引語の抽出法を提案する。**4.** では例として、土木関連用語辞典と文献の著者キーワードへの予備的な適用結果を示す。最後に **5.** で今後の課題を述べる。

2 関連研究

2.1 統計尺度を用いた複合語単位の抽出

文献[3]では、検索キーワードの利用者への提示を目的としたWeb文書からの用語獲得法を提案している。この方法では、任意の単語（形態素）nグラムについて、頻度やnグラム長に基づく統計スコアを計算する。そして、その値が極大となる（すなわち語頭・語尾の語を削除して得られる構成語よりも高いスコアを持つ）単語nグラムを特徴的な語として出力する。任意の単語列を対象とするため、複合名詞にとどまらず、様々な単位の用語の抽出が可能である。文献中ではまた、接辞からはじまる語を取り除くなど、表層的な処理の適用が併せて必要であると報告されている。

文献[4]では、複合語内の係り受け関係を有向グラフで表現して、HITSやPageRankの尺度を適用することにより重要度を計算する方法を提案している。実験では、先行語がつねに直後の後続語を修飾するという単純化した仮定を用いているが、複合語の内部構造に着目している点が特徴的である。

2.2 複合化規則に基づく用語抽出

上記が統計的な手法を適用しているのに対して、文献[5]では、パターンに基づく用語抽出モデルを提案している。具体的には、形態素の品詞に和語・漢語・カタカナ語の語種情報を加えた新たな文法カテゴリを設定し、これに基づき、(i) 形態素から単語、(ii) 単語から複合語、(iii) 複合語から名詞句、の3つのレベルの複合化パターンを記述している。テキストから用語候補を抽出する実践の場では、多くの場合「名詞の並びは複合名詞」のような単純なルールを適用していると考えられるが、文献[5]では再現率の観点から、詳細な接続関係を考慮したパターン記述が重要であることを指摘している。

2.3 形態素解析の拡張としての分かち書き処理

文献[6]では、入力されたテキストを形態素の頻度（重み）付き集合に変換するための分かち書きの一般化を提案している。従来の形態素解析による分かち書きが、接続コストを最小とする最適解だけを出力するのに対して、文献[6]の提案による一般化された分かち書きでは、あらゆる接続の可能性を考慮した期待値を出力する。文献中の例をとると、たとえば「京都大学」は、「京都大学」1語ではなく、重みを付与された語集合「京 (0.0014)」「京都 (0.3514)」「京都大 (0.1440)」「京都大学 (0.5032)」「都 (0.0014)」「大 (0.0255)」「大学 (0.3272)」「学 (0.1695)」と変換されることになる。可能性の低い単位には低い確率が割り当てられるため、ノイズの影響が抑えられる。この手法において、長い単語と短い単語（文字）のいずれを重視するかはパラメタで制御できるため、アプリケーション側で目的に応じた単位の設定が行えることが利点としてあげられている。

[2] http://www.jsce.or.jp/library/
[3] http://geta.ex.nii.ac.jp/

2.4 複合語の辞書登録

文献 [7] では、語単位の複数レベルにわたる情報を明示的に記述するために、複合語の構成語の依存構造を二分木で表現して関係データベースに格納する方法を用いている。文献 [8] における UniDic 電子辞書の設計では、語の単位として設定する第 0 層～第 3 層のうち、複合語および名詞的表現に対応する第 2 層・第 3 層について、その語構成を直接辞書に記入できるようになっている。また、文献 [1] では、形態素解析ツールの辞書に登録し、見出し語が複合語として分割可能であるかどうかのフラグを登録し、形態素解析の実行時にこれを参照して複合語出力の有無を切り替える 2 層の処理を提案している。

境界認定による日本語解析

文献 [9][10] では、語ではなく語境界とその種別を認定する「境界認定」を、形態素解析に代わる新しい日本語解析の枠組みとして提案している。また、このような境界認定システムの実現法の例として、形態素の連接行列の各要素に境界の種別を示す「境界 ID」を対応させて、実行時に参照する方法を示している。たとえば複合語の構造解析について文献中の例をとると、「ワイン城完成記念パーティ」の認識結果は 3 つの語境界、(i)「ワイン城」と「完成」を結ぶ「その他名詞と動詞性名詞の間の格（が）関係」、(ii)「完成」と「記念」を結ぶ「動詞性名詞と動詞性名詞の間の格（することを）関係」(iii)「記念」と「パーティ」を結ぶ「動詞性名詞と普通名詞の間の連体修飾（する）関係」、と認識されることになる。

文献 [9][10] ではさらに、境界 ID を連接行列の要素ではなく、辞書に登録された各見出し語毎に定義する方法を提案しており、このような方法により連接条件を記述するための品詞細分類が不要になる可能性を指摘している。

2.5 本稿の目的とタスク設定

本稿の目的は文献 [3] の場合と同様に、専門性の高いテキストを対象とした索引語の抽出である[4]。手法の検討では、特に以下の 2 点に注意を払った。

(1) 分野特有の用語を多く含むテキストを対象とすることから例外的な用法に容易に対応できること

(2) 言語に関する詳細な知識を持たない一般の利用者が簡単に辞書を管理・維持できること

上記を満足するため提案手法では、図 1 に示すように、(A) テキストからの最長複合単位の抽出、(B) 複合語内の構成語の依存解析、の 2 つを独立なモジュールで実現することとした。本稿では、両者のうち (B) に焦点をあてて検討を行う。

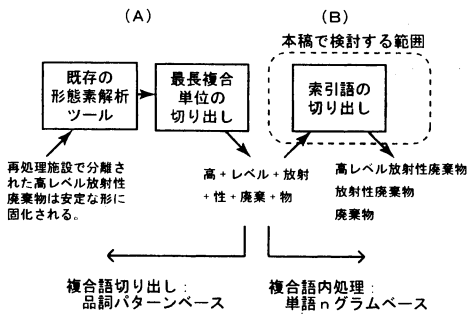


図 1: 想定する索引語抽出処理の流れ

(A) と (B) を異なるタスクとして設定する理由の第一は、品詞パターンベースの処理から単語 n グラムベースの処理への切り替えを行うためである。従来研究の多くは、品詞あるいは字種のカテゴリに基づき接続コストや複合化規則を定めるものである。しかしながら、分野固有の用法では、品詞や字種のカテゴリだけでは十分でない場合も見受けられる。また、形態素解析の結果が必ずしも妥当なものではない場合についても考慮が必要である。構成語の連なりが 1 つのまとまりとなって、特定の働きをする場合もある（これらの具体的な例については **4.1** で紹介する）。そこで本稿では (B) において、品詞カテゴリではなく単語 n グラムに対して接続コストを割り当てる方法を検討する。

(A) と (B) を異なるタスクとして設定する理由の第二は、利用者側で準備する辞書資源の簡単化である。品詞タグのついた学習用コーパスの準備には手間がかかる。そこで本稿では、タグなしの単純な見出し語集合をコーパスとして利用することとし、見出し語の語頭・語尾に位置する語の左側・右側境界が自明であることを利用してコストの調整を行う方法を検討する。

[4]検索要求としても分野に特化したものを想定しているので、ここでは、索引語として抽出するべき複合語は、分野に特化した用語、いわゆる専門用語であると考えている。

まとめると、本稿で設定する索引語抽出タスクは以下の通りである。

○パラメタ調整に用いるコーパス
　特定分野にかかわる見出し語集合

○入力
　関連分野のテキストから抽出した複合語

○出力
　適切な単位に分かち書きされた索引語の集合

3 提案手法

3.1 基本方針

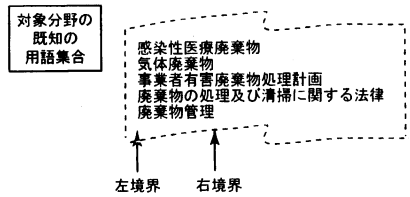
提案手法の基本方針は、(1) 対象分野の既知の用語（見出し語）集合を入力として構成語間の連接コスト計算のための重みを求め、(2) これに基づき与えられた任意の用語に対する依存木を導出して索引語の抽出を行う、ことである。ただし、ここで導出する複合語構成語の「依存木」は連接コストの重みから機械的に定まるもので、語彙概念構造に基づく語構成論の場合のような意味処理を意図したものではない。

(1) のコスト計算では、まず、与えられた見出し語集合を形態素解析ツールで分かち書きして、隣り合う形態素（以下、本稿では「単語」）間の連接コストの初期値を定める。次に、単語間の連接部分について、コスト最小のものから順位をつけ、図2のような依存木を求める。そして、導出した依存木の上で部分木をなす単語列を索引語候補として選択して連接コストの値を更新する。上記を一定回数反復して最終的な連接コストを決定する。(2) の索引語抽出では、同様にして作成した依存木から候補を数え上げ、後述する終端コストがある閾値以上のものを索引語として出力する。

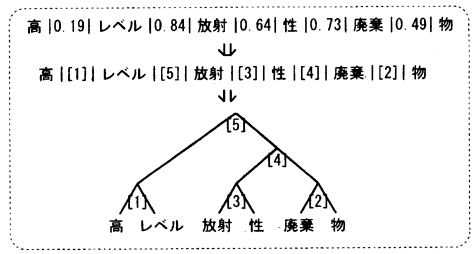
具体的には図2の場合、まず依存木の形から、「高レベル」「放射性」「廃棄物」「放射性廃棄物」「高レベル放射性廃棄物」の語が候補として選ばれる。次に、終端コストの制約から「高レベル」「放射性」が独立した用語としては適格でないと判断される。最終的に、「廃棄物」「放射性廃棄物」「高レベル放射性廃棄物」が残ることになる。

3.2 連接コスト計算のためのモデル

見出し語集合という限定された情報源を入力とする場合には、単語nグラム数の組み合わせごとに連接コストを設定すると接続テーブルのパラメタ数が



(a) 入力として利用する情報



(b) 解析結果

図 2: 複合語単位抽出手順の概要

膨大となり現実的ではない。そこで本稿では、「連接コストは前後の文脈に依存しない」という単純化をして、各単語nグラム（W で表記する）ごとに以下の確率を求める。

(a) 左方確率

　W を含む最小部分木上で W が左側に位置する確率（$P_{cl}(W)$）

(b) 右方確率

　W を含む最小部分木上で W が右側に位置する確率（$P_{cr}(W)$）

そして、各単語について、右連接コストを

$$Con_R(W) = -log(P_{cl}(W)) \quad (1)$$

左連接コストを

$$Con_L(W) = -log(P_{cr}(W)) \quad (2)$$

として、任意の単語nグラム W_i と W_j の間の連接コスト $Con(W_i, W_j)$ を以下で定義する。

$$\begin{aligned} Con(W_i, W_j) &= Con_R(W_i) + Con_L(W_j) \\ &= -log(P_{cr}(W_i)) - log(P_{cl}(W_j)) \end{aligned} \quad (3)$$

すなわち、連接コストは正の値で、ゼロに近いほど結びつきが強いとみなされる。たとえば図2-(b)の例では、連接コストは以下となっている。

$$\begin{aligned} Con(\text{高、レベル}) &= 0.19 \\ Con(\text{廃棄、物}) &= 0.49 \\ Con(\text{放射、性}) &= 0.64 \\ Con(\text{放射性、廃棄物}) &= 0.73 \\ Con(\text{高レベル、放射性廃棄物}) &= 0.84 \end{aligned} \quad (4)$$

依存木の導出では、単語nグラムを構成する部分列どうしの連接コストを計算し、その値が最小のものからボトムアップに部分木を構成する。連接コストの値が等しくなる場合に限り、依存木は3つ以上の枝分かれを持つことになる。ただしこれは表記上の都合であり、並置などのタイプに対応するものではない。

3.3 左方・右方確率の推定

連接コストの計算に必要な確率 P_{cl}、P_{cr} は以下のように定める。まず、与えら得た用語集合中での単語nグラム W について以下を数える。

(1) W が単独で見出し語となる回数 $freq(W)$
(2) W が用語の先頭に出現する回数 $freq^0(W\_)$
(3) W が用語の末尾に出現する回数 $freq^0(\_W)$
(4) W が用語の内部に出現する回数 $freq^0(\_W\_)$

用語集合中での W の総出現頻度は、$F(W) = freq(W) + freq^0(W\_) + freq^0(\_W) + freq^0(\_W\_)$ である。次に、左方確率および右方確率の初期値を以下で定める。

$$P^0_{cl}(W) = \frac{freq^0(W\_) + freq^0(\_W\_)}{F(W)} \quad (5)$$

$$P^0_{cr}(W) = \frac{freq^0(\_W) + freq^0(\_W\_)}{F(W)} \quad (6)$$

次に、上記を用いて用語集合中のすべての語について依存木を作成し、以降、単純な繰り返しアルゴリズムを適用して P_{cl} および P_{cr} の値を更新する。後述の実験では反復数20を用いている。

ここで、依存木を作成していない初期状態では、すべての可能な単語nグラムが出現したものとして頻度を数え上げるが、2回目以降では、導出した依存木の上で W を含む最小の部分木の中での出現位置に応じて $freq^t(W\_)$、$freq^t(\_W)$、$freq^t(\_W\_)$ を求める。すなわち、初期状態では「高レベル放射性廃棄物」に対して、「高」「高レベル」「高レベル放射」「高レベル放射性」「高レベル放射性廃棄」「高レベル放射性廃棄物」すべての $freq^0(W\_)$ の値に1を加え、2回目以降の図1の状態では「高レベル放射性廃棄物」に対して、「高」「放射」「廃棄」「放射性」「高レベル」の $freq^t(W\_)$ の値に1を加える。

3.4 適格な語の選別

単語nグラム W に対して、W が左側に語を伴う確率を左側非終端確率 $P_{bl}(W)$、W が右側に語を伴う確率を右側非終端確率 $P_{br}(W)$ として、以下で定義する。

$$P_{bl}(W) = P^0_{cr}(W) = \frac{freq^0(\_W) + freq^0(\_W\_)}{F(W)} \quad (7)$$

$$P_{br}(W) = P^0_{cl}(W) = \frac{freq^0(W\_) + freq^0(\_W\_)}{F(W)} \quad (8)$$

$P_{bl}(W)$ および $P_{br}(W)$ は便宜的に、各語が必須格を持つかどうかの目安として用いるものである。値の更新はしない。任意の W に対して $P_{br}(W)$ が1に近いとき、W の用法は接頭辞的であり、$P_{bl}(W)$ が1に近いとき接尾辞的であるといえる。そこで、これらの値が閾値（実験では0.95）より大きい場合には、W は独立の用語として適格ではないと判断することにする。たとえば上記で「高レベル」は右依存確率が1.0、すなわち与えられた用語集合内の用法では必ず右側に被修飾語を伴うので、索引語として選択されない。

3.5 未知語への対応

式(3)による連接コストは左右の語のコストの単純和で表現されるので、連接部分において一方の語が未知語である場合にも既知の語の情報を活用することができる。さて、上記では説明を簡単にするため、形態素解析で分かち書きされた語の表記を構成語としたが、実際には、「高：接頭辞-名詞接続」のように、前段で適用した形態素解析による品詞タグを付与した形で処理を行っている。そこで、品詞nグラムについても単語nグラムと同様に、連接コストや非終端コストを計算することにして、未登録語については、(i) 形態素解析の品詞情報が利用可能であるときは、その品詞カテゴリの平均値を、(ii) それ以外の場合には確率0.5をデフォルトの値として用いる。

品詞タグを付与することで、たとえば語尾の「法」（建築法）と語頭の「法」（法整備）はそれぞれ「名詞-接尾-一般」「名詞-一般」と区別され、間接的に形

態素解析ツールの品詞連接コストが考慮されることになる。しかしながら、用語単位で解析を行った場合の品詞と文中で出現した場合の品詞は必ずしも一致するわけではなく、品詞が誤って付与されることもある。したがって、品詞カテゴリに基づく情報はあくまで補助的なものと位置づけている。

4 土木関連用語集合への適用

4.1 土木学会用語集見出し語の分析

用語集合として土木用語大辞典 [11] の見出し語22,041語を用いて、提案手法を適用した。

まず、形態素解析ツール Chasen を用いてそれぞれの見出し語を分かち書きして、単語 n グラムに変換した。この場合の構成語の異なり数は 9,205 で、見出し語あたり平均構成語数は 2.55 であった。見出し語としては、2 語から構成されるものが最も多く全体の 44%、3 語以上のものは全体の 43%、構成語数の最多は「地方拠点都市地域の整備及び産業業務施設の再配置の促進に関する法律」の 17 語であった。また、他の見出し語と共通の構成語が 1 つもない見出し語は 92 個で全体の 高々0.4%であった。このことから、新規の用語であっても分野が同じであれば、構成語として既知の語が含まれる場合が多いことが予想される。

次に、分かち書きした見出し語集合をサフィックスアレイに変換して、すべての単語 n グラムについて連接コスト計算に必要な頻度情報 ($freq(W\_)$、$freq(\_W)$、$freq(\_W\_)$) を求めた。全体の単語 n グラム異なり数は 54,724 で、見出し語でないものが約 50%を占める。出現回数が最も多い構成語は「法:名詞-接尾-一般」の 622 回、以下「の:助詞-連体化」617 回、「式:名詞-接尾-一般」362 回、「工法:名詞-一般」336 回などであった。

出現回数が上位の単語に関する統計量を表 1 に示す。品詞タグは Chasen による。先頭に'*' がついている語は、用語辞典の見出し語としては出現せず、構成語としてのみ観察されたことを示す。単語ユニグラムの結果から、たとえば表記が「法」で品詞タグが「名詞-接尾-一般」である語は、語頭には一度も出現せず、622 回中語尾での出現が 604 回、残りの 18 回は語中で前後に語を伴って出現した等がわかる。このように、形態素解析ツールによる品詞タグは構成語の出現位置と密接に関係している。しかしながら、表 1 によると、必ずしも品詞タグが分野での用法に一致するわけではない。たとえば「法」と「式」は同じ「名詞-接尾-一般」であるが、「法」は「乱数発生法」「連続給砂法」のように 97%のケースで語尾に出現するのに対して、「式」は「揚水式発電」「連続式掘削機」のように、前後に語を伴って出現するケースが 82% を占める。また、「工法」は「名詞- 一般」とタグ付けされているが、分野での用法はすべて語尾表現である。「工法」は単独では用語とみなされていないことから、用法の上では接尾時と同様であると考えられる。

一方、表 1 で構成語数 2 以上の頻出語の中には、「骨 | 材」「支 | 保 | 工」「トラ | ス」「灌 | 漑」などの解析誤り、「二 | 次」「一 | 次」のようにまとまりを持って語頭に出現するもの、「の | 原理」のように語尾に出現するもの等さまざまなパターンが見られる。

4.2 用語抽出例

次に、前節で述べた方法により求めた連接コストおよび非終端コストを索引語の抽出に適用した例を示す。

実験では、土木関連文献 53,894 件の著者キーワードにおける頻出語のうち、構成語数が 3 語、および 4 語以上の 2 つの場合について、それぞれ出現頻度が上位の 100 件ずつを選び入力とした。そして各語に対して依存木を導出し、左右の非終端確率がいずれも 0.95 以下のものを索引語として抽出した。何を索引語の正解とみなせばよいかは議論を要する問題であるが、ここでは土木用語大辞典の見出し語に含まれている索引語が抽出できたかどうかを判断の基準とした。

抽出された索引語の例を図 3 に示す。図 3-(a) は、抽出した依存木が正しく、索引語としても適切なものが選択された場合である。図 3-(b) は、依存木は必ずしも妥当とはいえないが、不適切な単語列（この場合は「的非線形」）が非終端コストに関する条件から取り除かれ、索引語としては適切なものが選択された場合である。図 3-(c) は、依存木の誤りが原因で、想定した索引語が選択されなかった場合である。本稿で想定している枠組みでは、(a) および (b) の場合は正解、(c) の「土圧」「土圧係数」が不正解となる。ここで、入力語の係り受け関係には必ずしも一定の規則があるわけではない。たとえば、構成語数が 3 語の場合であっても、依存木が左分岐型になるものと右分岐型になるものはいずれも相当数見受けられた。

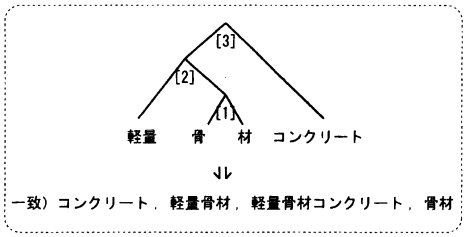
表 1: 頻度上位の単語に関する統計量

単語ユニグラム					
$F(W)$	$freq^0(W\_)$	$freq^0(\_W)$	$freq^0(\_W\_)$	語	品詞
622	0 (0.00)	604 (0.97)	18 (0.03)	*法	(名詞-接尾-一般)
617	0 (0.00)	0 (0.00)	617 (1.00)	*の	(助詞-連体化)
362	0 (0.00)	66 (0.18)	296 (0.82)	*式	(名詞-接尾-一般)
336	0 (0.00)	336 (1.00)	0 (0.00)	*工法	(名詞-一般)
319	72 (0.23)	207 (0.65)	40 (0.13)	*計画	(名詞-サ変接続)
319	112 (0.35)	59 (0.18)	148 (0.46)	*水	(名詞-一般)
264	0 (0.00)	225 (0.85)	39 (0.15)	*量	(名詞-接尾-一般)
264	0 (0.00)	108 (0.41)	156 (0.59)	*性	(名詞-接尾-一般)
217	54 (0.25)	105 (0.48)	57 (0.26)	波	(名詞-一般)
213	55 (0.26)	112 (0.53)	45 (0.21)	コンクリート	(名詞-一般)
200	0 (0.00)	195 (0.97)	5 (0.03)	*係数	(名詞-一般)
190	0 (0.00)	174 (0.92)	16 (0.08)	*率	(名詞-接尾-一般)
188	7 (0.04)	170 (0.90)	11 (0.06)	*試験	(名詞-サ変接続)
186	0 (0.00)	12 (0.06)	174 (0.94)	*型	(名詞-接尾-一般)
178	0 (0.00)	97 (0.54)	81 (0.46)	*流	(名詞-接尾-一般)
174	28 (0.16)	58 (0.33)	88 (0.51)	*圧	(未知語)
173	75 (0.43)	25 (0.14)	72 (0.42)	交通	(名詞-一般)
172	34 (0.20)	74 (0.43)	64 (0.37)	*構造	(名詞-一般)
172	17 (0.10)	125 (0.73)	30 (0.17)	*事業	(名詞-一般)
172	0 (0.00)	157 (0.91)	15 (0.09)	*方式	(名詞-一般)
単語バイグラム					
$F(W)$	$freq^0(W\_)$	$freq^0(\_W)$	$freq^0(\_W\_)$	語	品詞
52	10 (0.19)	29 (0.56)	12 (0.23)	骨 材	(名詞-一般)(名詞-接尾-一般)
44	10 (0.23)	27 (0.61)	6 (0.14)	座 屈	(名詞-一般)(未知語)
42	3 (0.07)	33 (0.79)	6 (0.14)	*構造 物	(名詞-一般)(名詞-接尾-一般)
42	11 (0.26)	19 (0.45)	11 (0.26)	廃棄 物	(名詞-サ変接続)(名詞-接尾-一般)
40	33 (0.82)	0 (0.00)	7 (0.17)	*二 次	(名詞-数)(名詞-接尾-助数詞)
37	12 (0.32)	10 (0.27)	14 (0.38)	圧 密	(未知語)(名詞-一般)
35	5 (0.14)	27 (0.77)	2 (0.06)	交通 量	(名詞-一般)(名詞-接尾-一般)
35	0 (0.00)	32 (0.91)	3 (0.09)	*整備 事業	(名詞-サ変接続)(名詞-一般)
31	24 (0.77)	0 (0.00)	7 (0.23)	*一 次	(名詞-数)(名詞-接尾-助数詞)
27	6 (0.22)	19 (0.70)	1 (0.04)	灌 漑	(未知語)(未知語)
27	3 (0.11)	0 (0.00)	24 (0.89)	*支 保	(未知語)(名詞-一般)
27	23 (0.85)	0 (0.00)	4 (0.15)	*波 の	(名詞-一般)(助詞-連体化)
25	6 (0.24)	14 (0.56)	4 (0.16)	土 圧	(名詞-一般)(未知語)
25	5 (0.20)	16 (0.64)	3 (0.12)	トラ ス	(名詞-一般)(動詞-自立)
25	1 (0.04)	23 (0.92)	0 (0.00)	支 保 工	(未知語)(名詞-一般)(名詞-一般)
25	0 (0.00)	24 (0.96)	1 (0.04)	*保 工	(名詞-一般)(名詞-一般)
24	2 (0.08)	21 (0.88)	0 (0.00)	支 承	(未知語)(動詞-自立)
24	0 (0.00)	24 (1.00)	0 (0.00)	*の 原理	(助詞-連体化)(名詞-一般)
23	0 (0.00)	23 (1.00)	0 (0.00)	*措置 法	(名詞-サ変接続)(名詞-接尾-一般)
22	0 (0.00)	20 (0.91)	1 (0.05)	沈殿 池	(名詞-サ変接続)(名詞-一般)

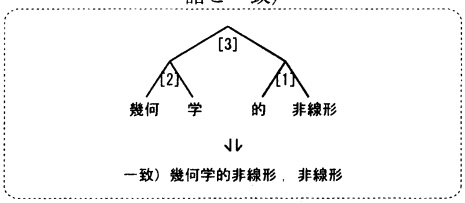
入力語の構成語数が 3 語、および 4 語以上の 2 者について、入力語から抽出した索引語と用語辞典の見出し語との一致数を調べた結果を表 2 に示す。入力語が、用語辞典の見出し語としてすでに登録されている場合とされていない場合のそれぞれについて数値を示してある。抽出に失敗した例の多くは、「支 | 承」の「承」や「ガ | スト」の「スト」のように、未知語を構成する語の一部（カタカナ語と漢字 1 文字）が独立な用語として抽出されてしまったケースであった。これらのうち高頻度で観察されるものは、形態素解析ツールの辞書登録語の候補となると考えられる。

5 おわりに

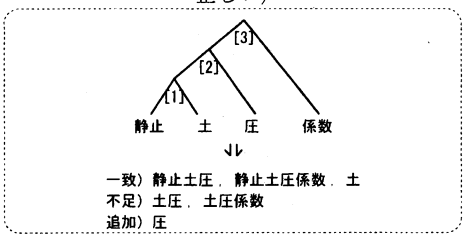
本稿では、専門的なポータルサイトにおける複合語キーワード抽出の問題に焦点をあて、分かち書き単位が層構造となる場合の解析法について、現在検討中の手法と予備的な実験結果を示した。提案手法では、専門分野における固有の用法に対応すること、および言語処理の専門家でない管理者による運用を容易にすることを特に意識して、品詞パターンベースと単語 n グラムベースの処理を切り分け、本稿では後者について検討した。本稿の予備実験では用語辞典の見出し語との一致数を調べたが、実際には用語辞典だけに含まれる語の中にも「不飽和」のよう



(a) 成功した例（抽出した用語が用語辞典の見出し語と一致）



(b) 成功した例（解析誤りはあるが抽出した用語は正しい）



(c) 失敗した例（解析誤りによる抽出もれ）

図 3: 解析の例

に接頭辞的なものが含まれたり、著者キーワード中に用語辞典には含まれていないが分野の用語とみなせるものが含まれたりする場合があり、必ずしも見出し語は基準として適切ではない。今後は手法の詳細化を行うとともに評価手法の検討を進める予定である。

表 2: 論文著者キーワード頻出語の解析結果

入力語の構成語数が 3 の場合（100 語中）	
入力語が用語辞書見出し語と一致	67 語
抽出した索引語が	
用語辞書見出し語と共通	141 語
用語辞書に含まれるが抽出に失敗	12 語
抽出されたが用語辞典に含まれない	45 語
入力語が用語辞書見出し語と一致	33 語
抽出した索引語が	
用語辞書見出し語と共通	42 語
用語辞書に含まれるが抽出に失敗	3 語
抽出されたが用語辞典に含まれない	56 語

入力語の構成語数が 4 以上の場合（100 語中）	
入力語が用語辞書見出し語と一致	48 語
抽出した索引語が	
用語辞書見出し語と共通	123 語
用語辞書に含まれるが抽出に失敗	15 語
抽出されたが用語辞典に含まれない	38 語
入力語が用語辞書見出し語と一致	52 語
抽出した索引語が	
用語辞書見出し語と共通	69 語
用語辞書に含まれるが抽出に失敗	23 語
抽出されたが用語辞典に含まれない	142 語

参考文献

[1] 青木和夫、中山章弘、松崎剛士「形態素解析での効率的な複合語処理」情報処理学会研究報告. NL, 自然言語処理 2003(57), 1-6, 2003.

[2] 野末道子、相澤彰子、坂本真至、今尚之、藤澤泰雄、小松淳、吉崎保、江口知秀「土木学会図書館書誌目録データベース検索次世代型インタフェースの構築と評価」第 27 回ディジタル図書館ワークショップ, 2005.

[3] 山本英子、池野篤司、濱口佳孝、井佐原均「検索支援に向けた Web 文書集合からの用語獲得」情報処理学会研究報告. NL, 自然言語処理 2004(108), 171-176, 2004.

[4] 辻河亨、吉田稔、中川裕志「語彙空間の構造に基づく専門用語抽出」情報処理学会研究報告. NL, 自然言語処理 2004(1),155-162, 2004.

[5] 竹内孔一, 影浦峡, ダイユベアトリス, 小山照夫「多言語専門用語抽出モデルの構築」言語処理学会第 11 回年次大会発表論文集, 887-790, 2005.

[6] 工藤拓「形態素周辺確率を用いた分かち書きの一般化とその応用」言語処理学会第 11 回年次大会発表論文集, pp.592-595, 2005.

[7] 浅原 正幸、米田 隆一、山下 亜希子、伝 康晴、松本 裕治「語長変換を考慮したコーパス管理システム」情報処理学会論文誌 Vol.43 No.07, pp.2091-2097 2002.

[8] 伝康晴「話し言葉研究に適した電子化辞書の設計」第 2 回話し言葉の科学と工学ワークショップ講演予稿集 39-46, 2002.

[9] 佐藤理史「境界認定の提案：(1) コンセプトと実現法 (解析)」情報処理学会研究報告. NL, 自然言語処理 2004(108), 25-32, 2004.

[10] 佐藤理史「境界認定の提案：(2) 背景と思想 (解析)」情報処理学会研究報告. NL, 自然言語処理 2004(108), 33-40, 2004.

[11] 土木学会編「土木用語大辞典」技報堂出版, 1999.