



Title	述語項構造解析による検索対象細分化を用いた自然文検索
Author(s)	山田, 浩之; ジェプカ, ラファウ; 荒木, 健治
Citation	電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, 110(400), 59-64
Issue Date	2011-01-20
Doc URL	https://hdl.handle.net/2115/63663
Rights	Copyright©2011 IEICE
Type	journal article
File Information	110008676788.pdf



述語項構造解析による検索対象細分化を用いた自然文検索

山田 浩之[†] ジェプカ ラファウ[†] 荒木 健治[†]

[†] 北海道大学大学院情報科学研究科 〒060-0814 北海道札幌市北区北 14 条西 9 丁目

E-mail: [†] {yamada_h, kabura, araki}@media.eng.hokudai.ac.jp

あらまし 本手法では、述語項構造解析を用いて、自然言語文で表現されたクエリに含まれる単語、係り受け関係に対して、項構造によるスコア付与を行い、データベースからの検索を行った。特に、述語項構造解析によって「ガ格」として判別された単語に対して、検索文章中で主となる「検索主体」を定義することによって検索を行う。これにより、単語によるスコアの偏りによる問題が是正できる。述語項構造解析には SynCha を用い、実験では交通事故の判例、「教えて!goo」のテキストを対象とした。

キーワード 自然文クエリ、情報検索、判例検索

Natural Language Search Based on Subdividing Objects of Search Using Predicate Argument Structure Analysis

Hiroyuki YAMADA[†] Rafal Rzepka[†] and Kenji ARAKI[†]

[†] Graduate School of Information Science and Technology, Hokkaido University

Kita 14 Jo Nishi 9 Chome, Kita-ku, Sapporo-shi Hokkaido, 060-0814 Japan

E-mail: [†] {yamada_h, kabura, araki}@media.eng.hokudai.ac.jp

Abstract In this paper we describe our system for retrieving texts using predicate argument structure analysis for queries in natural language. A dependency relation included in a query decides a word score, which is retrieved by defining "retrieval subject" that further becomes the main clue for retrieving sentences using a word classified by predicate argument structure analysis tool SynCha as "-ga case". As a result, the problem of biased word score can be corrected. In this research we target judicial precedents of traffic accidents and use "Oshiete! goo" online QA community entries for experiments.

Keyword Natural language query, Information Retrieval, judicial precedents extraction

1. はじめに

現在、Web 上では教えて!goo[1]や Yahoo!知恵袋[2]等の Q&A コンテンツが盛んに利用されている。これらのサイトでは、ユーザが質問したい内容について自然言語文で記述をし、回答者である他のユーザの助言を待つという形で運営されている。しかし、中には回答が得られないものや、高度に専門的な知識を要する質問、また回答者に信頼性をおけないことがある等の問題がある。そういった場合、情報検索等に精通していないユーザにとっては、必要とする情報を得ることは容易ではない。特に、裁判の判例に関する質問においてそのような問題が顕著である。

判例とは、過去の裁判の内容を文書としてまとめたものである。判例を参照することにより、今後行われる裁判に対してどのような判決が導かれるかということをも大まかに推測する事が可能となる。そのため、発生した事案に対して類似する過去の判例を検索する事は大変有益な事である。

特に、交通事故に関する判例は交通事故自体がとても身近な事件であり、一般の人々にとっても有益であ

る。具体的に有益な場面としては、交通事故発生後の保険会社との損害額に関する交渉を行う際の妥当性の検証や、請求額の根拠としてなどである。

既存の判例検索システム[3][4][5]では、索引語を用いたキーワード検索が主流である。しかし、キーワードだけの検索では要求した文と合致した検索結果を得ることは難しい。例えば「右折」や「衝突」といった単語を検索のクエリとして用いたとする。その場合、「衝突」という単語は今回実験の対象とした判例 1,438 件の判例の内、1,295 件の判例に出現するため判例を絞り込むための有効なクエリとはならない。また、「右折」という単語による検索結果を見てみると、「右折する」という行為以外でも「右折専用レーン」や「右折禁止」、「右折青矢印」等の部分でも「右折」という単語が用いられているため、そのままクエリを用いたところで状況に関する絞込みは望めない。

そのため、自然言語文を入力として利用することで、単語のみではなく、係り受けなどの周辺情報を利用することができると考える。

このような点から、本研究では Q&A コンテンツで

自然言語文として投稿される文を入力として、入力された文と最も合致した文章を検索するための手法についての提案を行う。

本手法では、文の解析には SynCha[6][7][8][9][10]を用いる。SynCha は奈良先端科学技術大学院大学情報科学研究科、自然言語処理学講座(松本研究室)にて開発された、述語項構造解析を行なうプログラムである。形態素解析・係り受け解析などの結果を受け取り、「誰が」「何を」「何に」「どうした」といった情報を出力することができる。

処理過程においては、入力文中で SynCha による解析で述語や他の格と大きく関わりのある「ガ」格の単語を「検索主体」として定義し、検索においてさらなるスコア付与を行う。

実験では、まず一般的な文書を対象として、「教えて!goo」からランダムに選択した「社会・職場」カテゴリと「家事・生活」カテゴリに対するシステムの適用・評価を行った。次に、専門的な文書である「交通事故判例」に対しての実験を同様に行った。

評価は、単語のみの検索、単語と係り受けを考慮した検索をベースとして、平均順位・平均逆順位を尺度として行った。

2. 関連研究

自然言語文での検索に関する研究としては、新里ら[11]の研究がある。新里らの研究では、自然言語で表現されたクエリに含まれる単語、係り受け関係を重要度に応じて使い分けて文書検索を行う方法を提案している。しかし、入力文には端的に表現された自然文クエリが使われるなど、より複雑な自然文クエリを扱っていない。

また、他の自然言語文の入力による検索の研究としては黒橋ら[12]の研究が挙げられる。この研究では、知識ベースを自然言語によって記述し、ユーザの入力質問と知識ベースとの柔軟なマッチングを行うことを処理の基本としている。しかし、この手法では単文による入力しか想定されておらず、またあらかじめ自然言語によって記述された検索対象が用意されているため構文木によるマッチングが容易になっている。

判例検索に関する研究としては、原田らの研究が挙げられる。原田ら[13]の研究では、意味解析に SAGE[14][15]、文脈解析に InSeRA[16]を用いて、文章が表す内容を語が格納されたノードと語間の関係を表すアークからなる意味グラフとして捉え、判例文と問い合わせ文の文章間の内容類似度を算出することで検索を行っている。

しかし、これらのシステムでは入力時に詳細な事故

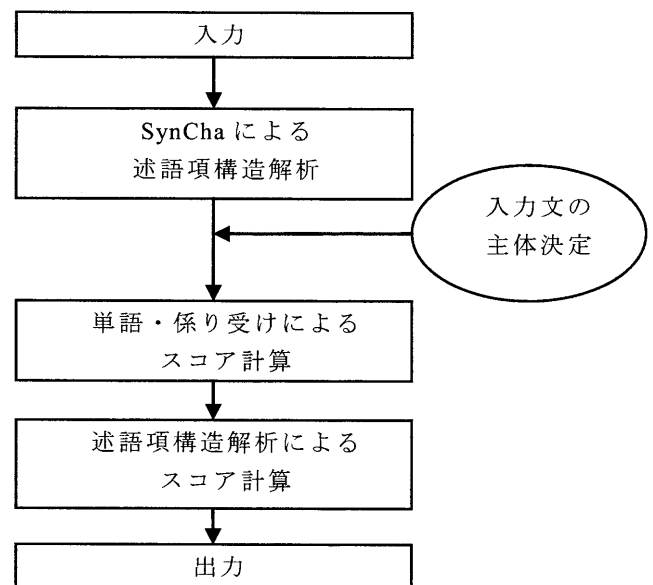


図 1: 処理過程

の状況を記述する必要があり、ユーザにとっての負担は小さくない。

その他の判例に関する研究としては、江越ら[17]の研究がある。江越らの研究では判例が一つの文書ではなく、役割の異なる複数の文書から構成されていることに着目し、語の偏りによる $tf \cdot idf$ 値の修正等が行われている。しかし、この手法では交通事故判例のような、判例のどの部分に対しても「右折」や「直進」といった事故に関する単語が出現する文書においては語の偏りによって事故の状況を特定するのは困難であると考えられる。また、法律に対しての知識が無いユーザの入力には、判例内で使用される語と同様の語が必ずしも用いられるわけではなく、そういった判例内にはあまり登場しない語を検索対象の判例の単語としかしてマッチングさせるかという課題に対する考慮がなされていない。

以上の自然言語文での検索による研究と判例検索に関する研究を考慮し、本研究では下記に挙げる条件を満たすシステムの構築を行う。

1. 特別な知識を必要としない自然文による入力を対象とする
2. タスクを限定せず、検索対象を一般的な文書から専門的な文書まで幅広いものを対象とする

3. システムのアルゴリズム

3.1. 概要

システムの処理過程を図 1 に示す。

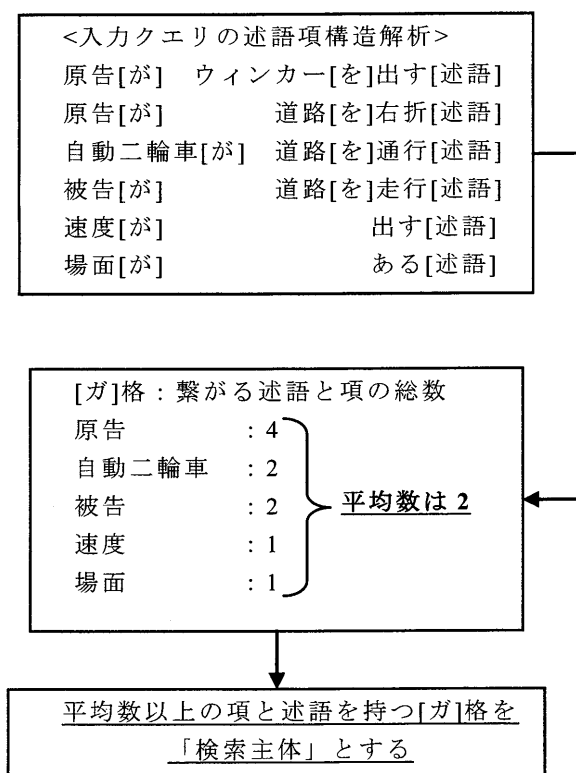


図 2:入力文の主体決定

3.2.入力

自然言語文による入力を行う。

なお、今回の交通事故判例検索実験においては、人手でクエリの拡張を行った。

3.3.SynCha による述語項構造解析

入力された文に対して、SynCha を用いて形態素解析、係り受け解析、述語項構造解析を行う。

3.4.入力文の検索主体決定

入力文に対して、SynCha による述語項構造解析を施し、検索において注目すべき「検索主体」となる語を決定する。これは、情報検索において、話題の中心となる語が文章中に含まれており、その語には検索時に特別なスコアを与えることにより、通常の単語・係り受け解析よりもより精度の高い検索を目指す。

検索主体の決定方法を、以下に示す。

1. 入力された検索クエリ文に対して、SynCha による述語項構造解析を施す。
2. 「ガ」格に対して、付与された「二」格、「ヲ」格、「述語」をカウントする。
3. 「ガ」格に対して、付与された「二」格、「ヲ」格、「述語」の総数をカウントする。

$$R(q, d) = (1 - \beta) \sum_{t \in T_{qword}} BM_{25}(t, d) + \beta \sum_{t \in T_{qpnd}} BM_{25}(t, d)$$

T_{qword} :クエリから抽出された単語の集合

T_{qpnd} :係り受けの集合

β :スコアリングに係り受けをどの程度

利用するかのパラメータ（本手法では 0.2）

図 3:OkapiBM25 計算式

$$BM_{25}(t, d) = w \times \frac{(k_1 + 1)F_{dt}}{K + F_{dt}} \times \frac{(k_3 + 1)F_{qt}}{k_3 + F_{qt}}$$

$$w = \log \frac{N - n + 0.5}{n + 0.5} \quad K = k_1 \left((1 - b) + b \frac{l_d}{l_{ave}} \right)$$

F_{dt} :文書 d 中での t の出現頻度

F_{qt} : q 中での t の出現頻度

N :検索対象となっている文書数

n : q の文書頻度

l_d : d の文書長（単語数）

l_{ave} :平均文書長

k_1, k_2, b は Okapi のパラメータであり、

$k_1=1, k_2=0, b=0.6$

図 4:OkapiBM25 のパラメータ

4. 「二」格、「ヲ」格、述語の総数を、「ガ」格の異なり数で割り、算出された平均値を検索主体決定の閾値とする。
5. 検索主体決定の閾値を上回る、「二」格・「ヲ」格・「述語」を付与された「ガ」格を「検索主体」として決定する。
6. 以上のように決定された「検索主体」をスコア計算時に考慮する。

なお、上記の「検索主体」算出方法の例を図 2 に示す。

3.5.単語・係り受けによるスコア計算

検索対象とするデータベースに対して、あらかじめ名詞・動詞の単語スコア、係り受けに対するスコアを付与しておき、入力文とのマッチングによってスコアを決定する。

なお、スコアの計算には新里らの手法において用いられた、単語と係り受けのスコアを計算できるように拡張された OkapiBM25[18]を用いた。

また、検索対象の文書に対しても SynCha による述語項構造解析を施し、それぞれの格と述語に対しても、同様に OkapiBM25 によるスコア付けを行う。

OkapiBM25 の式については、図 3 および図 4 に示す。

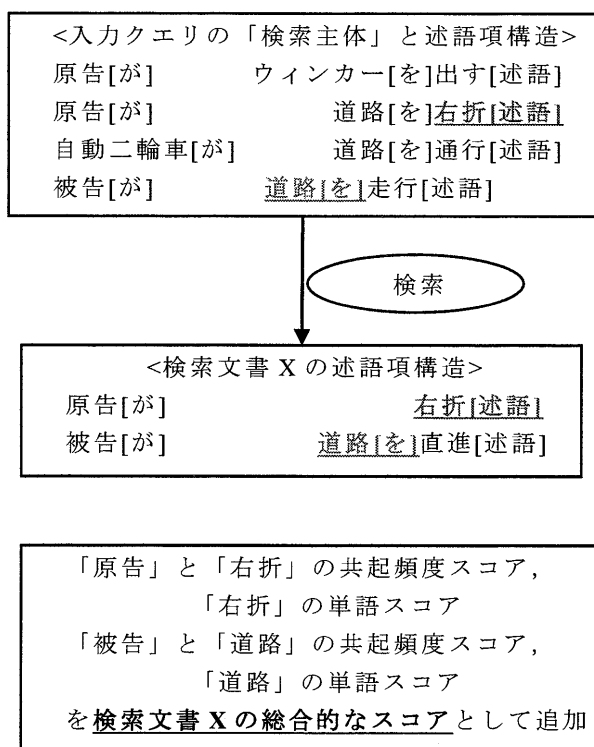


図 5: 検索主体を用いたスコア計算例

3.6. 述語項構造解析によるスコア計算

検索対象とするデータベースに対して、SynCha による述語項構造解析を行い、その結果に対してスコア付けを行う。SynCha によって、「ガ」格、「ニ」格、「ヲ」格、述語として判断された形態素に対して、OkapiBM25 によるスコア計算を行う。

この時、3.4 において「検索主体」として判断されたものに対して、「ガ」格とその項のマッチングによって、「ガ」格との共起頻度、その項の出現頻度を付与する。

なお、計算手法の例については、図 5 および図 6 に示す。

3.7. 出力

計算されたスコアを総合的に高い値となった順にユーザに提示する。

4. 評価実験

4.1. 実験概要

検索実験は、一般的な文書と専門的な文書の 2 種類に対して行った。

本稿において、一般的な文書は、特段の専門的な知識を有しないユーザ同士で質問・解答を行えるものと考え、そのように判断できる「教えて!goo」のカテゴリ内からランダムに「社会・職場」、「家事・生活」を選択した。また、専門的な文書としては解答時に専門的な知識を有するものとして、「交通事故判例」を対象

文書 X のスコア = OkapiBM25 による
単語と係り受けのスコア
+ 「ガ」格によるスコア

「ガ」格によるスコア
$$= \alpha \times (\Sigma \text{ ヒットした検索主体と格の共起頻度スコア} + \Sigma \text{ 格の単語スコア})$$

$$\alpha = \frac{\text{クエリにおける検索主体と共起する述語と格の総数}}{\text{文書 X における検索主体と共起する述語と格の総数}}$$

図 6: 「検索主体」のスコアへの換算法

とした。

また、一般的な文書を対象とした実験においては、「教えて!goo」内のユーザの質問文を「入力文」、他ユーザからの回答文を「正解」としてテストセットの作成を行った。これは、同様の話題について話されているため、検索を行う際には類似の文として扱えると考えたためである。

一般的な文書を対象とするものとして、①「教えて!goo」の「社会・職場」カテゴリ内からランダムで抽出した 1,006 件に対して、その質問文を入力とし、回答文を正解として検索を行った。②同様に、「教えて!goo」の「家事・生活」カテゴリから 304 件の質問文・回答文を抽出し、同様の実験を行う。

また、専門的な文書の対象として、③教えて!goo の「交通事故」カテゴリ内から、回答文に正解となる判例が明確に記述されているものを 12 件選び、その質問文を入力として正解となる判例を LEX/DB インターネット[19]から抽出した 1,438 件の判例の中から検索する。また、「交通事故」カテゴリの入力文は、判例に適合させるために人手でクエリを拡張した。

また、名詞・動詞のスコア付けには OkapiBM25 を用い、クエリの形態素・係り受け解析・述語項構造解析には SynCha を用いた。

比較実験は、

- ① 検索クエリの名詞・動詞のスコアのみを用いたもの
- ② 検索クエリの名詞・動詞のスコアかつ係り受けを利用した新里らの手法
- ③ 本手法

の 3 つの手法を比較した。

なお、評価式は、平均順位と平均逆順位を用いた。平均順位は複数ある正解として設定した判例全ての出力された順位を平均したものである。平均逆順位は、最も上位に出力された正解の順位の逆数を取り、平均したものであり、高い値を示すほどシステムとして正

表 1:「社会・職場」における実験結果

	平均逆順位	平均順位
① 単語のみ	0.081	433.1
② 単語＋係り受け	0.089	403.2
③ 本手法	<u>0.097</u>	<u>384.3</u>

表 2:「家事・生活」における実験結果

	平均逆順位	平均順位
① 単語のみ	0.104	390.2
② 単語＋係り受け	0.112	372.6
③ 本手法	<u>0.160</u>	<u>331.8</u>

表 3:交通事故判例における実験結果

	平均逆順位	平均順位
① 単語のみ	0.072	481.2
② 単語＋係り受け	0.143	442.4
③ 本手法	<u>0.253</u>	<u>282.6</u>

解文書を上位に出力しやすいという尺度として用いられる。

4.2.実験に用いた判例データ

抽出された 1,438 件の判例は、「交通事故判例検索」に含まれる判例のうち、死亡事故を含まず、事故に対する明確な答えとなる「過失割合」を含むものとする。

過失割合とは、交通事故におけるお互いの過失(不注意)の度合いを割合で表したものである。そのため、類似した事案に関しては近い割合の過失割合が導き出される。この過失割合を含むものを検索対象とした理由としては、入力文に対しての答えとなる判例を設定する際の手がかりとするためである。また、今回死亡事故を対象外とした理由は、Q&A サイトにおける死亡事故に関する相談が極端に少なかったことに起因する。

5. 実験結果

実験の結果を表 1,表 2,および表 3 に示す。

「社会・職場」においては、単語と係り受けを用いた手法と、本手法との間では平均逆順位で 0.008、平均順位で 18.9 ポイントの改善が見られた。

次に、「家事・生活」においては、単語と係り受けを用いた手法と、本手法との間では平均逆順位で 0.048、平均順位で 40.8 ポイントの改善が見られた。

そして、「交通事故判例」においては、単語と係り受けを用いた手法と、本手法との間では平均逆順位で 0.110、平均順位で 159.8 ポイントの改善が見られた。

いずれの場合においても、本手法がベースラインとなる手法を上回り、特に専門的文書として用いた「交通事故判例」に対して有効に働いたことが評価実験の結果から確認された。

6. 考察

実験結果において、3 つの実験に対してベースライ

ンよりも高い値を示したことから、本手法の有効性が示された。

また、交通事故判例において、最も他手法よりも改善が見られた。これは、判例の特徴上、

1. 主に、「被告」と「原告」の行動について記述されているため、「検索主体」としてスコア付けの際に他の文書よりも有効に働いた。
2. 判例は異なり語数が少ないため、正しい「検索主体」が選出された。
3. 判例の性質上、「これ」、「それ」などの曖昧な表現が文書中にあまり見られない

などといった理由が挙げられる。

交通事故判例については、特に上記の 1.に関して非常に有効であった。判例の性質上、原告と被告の行動に関して詳細に述べ、それに対して判決を下す。判例の文章上でも、「原告」と「被告」を指す言葉が変わることが少なく、「原告は」、「被告が」といった形で記述されていく。そのため、SynCha の「ガ」格として抽出されることが多く、またそれによってシステムが「検索主体」として選出され、検索時に有効に利用された。また、「原告」、「被告」以外にも、「自動車」、「自転車」等の事故の対象となる行動の主体も本システムで「検索主体」として選出されており、このこともシステムの結果に対して有効に働いていた。

これに対して、「教えて!goo」を対象とした実験においては、「検索主体」を適切に設定できない文書も多く見られた。これは、相談内容によってはかなり多くの種類の単語を用いており、主体の決定が困難なものがあつたためである。また、「教えて!goo」においては多数のユーザが利用するために、質問文・回答文として記述される文体も大きく異なっている。特に多く見られた検索が適切に行えなかった例としては、職場や家庭等の問題について詳細に記述するよりも、質問者の感情的な話題になり、検索主体として「私」という単語を選択してしまっているものなどである。現在のところでは、こういった人称代名詞に対して特別な処理を行っていないため検索を行う上での障害となってしまう。同様に、「それ」、「あれ」等の指示代名詞に対しても、対象となる部分を特定する処理を行っていないため、検索時には有効に働かなかった。人称代名詞、指示代名詞の問題を解決するためには、①指示代名詞の対象となる部分を特定する、②人称代名詞、指示代名詞が入力文内に多数出現する場合には、それらの語を除くなどの処理を行って、適切な「検索主体」を設定できるように閾値を文書ごとに柔軟に設定する、等の改善が必要であると考えられる。

他の適切な検索が行えなかった例としては、述語項構造解析の精度による問題と、「都会っ子」などの自然

文特有の固有表現への対応の問題、クエリの拡張を行っていないことによる単語のマッチング上の問題等が挙げられる。

また、判例に対しての実験では、例えば「右折車と直進車の事故」については直進車が右折車の前方から来たものか、後方から来たものかといった位置などに関する状況理解がなされていない。そういったことが、検索のエラーとして影響を与えていた。これらの問題については、①SynChaで対象としていない「の」、「から」、「へ」などの助詞の利用、②判例特有の表現の分析・利用をすることによって改善されると考えられる。

以上のことより、今後システムを改善していく上では、

1. 指示代名詞が対象とする部分の特定
2. 人称代名詞が多数出現した場合等に、「検索主体」を決定するための閾値を適切に設定する手法の確立
3. 述語項構造解析の改良
4. 適切なクエリ拡張の選択
5. SynChaで対象とされていない助詞の利用
6. 対象とする文書ごとの特徴の分析

が必要であると考えられる。

7. まとめ

本稿では、述語項構造解析を用いて、「検索主体」を設定することによって検索システムの精度向上を行った。特に、自然言語文を対象として、一般的な文書である「教えて!goo」の「社会・職場」カテゴリ、「家事・生活」カテゴリと、専門的な文書である「交通事故判例」という2つの大きく異なる文書を対象として性能評価実験を行った。

その結果、単語や係り受けを単純に利用したものよりも、本稿で示したシステムを利用することで精度が向上することが示された。実験においては、一般的な文書として用いた「教えて!goo」の「社会・職場」カテゴリにおいて、平均逆順位で 0.008、平均順位で 18.9 ポイント、「教えて!goo」の「家事・生活」カテゴリにおいて、平均逆順位で 0.048、平均順位で 40.8 ポイントの改善が見られ、専門的な文書を対象として行った、「教えて!goo」の「交通事故」カテゴリ内の質問文を入力として、「交通事故判例」を検索する実験においては、平均逆順位で 0.110、平均順位で 159.8 ポイントの改善が見られた。

今後の予定としては、指示代名詞の示す部分を特定する手法の利用、人称代名詞が入力文に出現した場合、その出現頻度等を考慮した上で、「検索主体」を決定するための閾値を適切に設定する手法の確立、SynChaで対象としていない助詞の利用等を行う必要があると考えられる。

文 献

- [1] 質問&回答 (Q&A) コミュニティ - 教えて goo, <http://oshiete.goo.ne.jp/>
- [2] Yahoo!知恵袋, <http://chiebukuro.yahoo.co.jp/>
- [3] レクシスネクシス・ジャパン, <http://www.lexisnexis.jp/>
- [4] LLI 判例検索システム, <http://www.lli-hanrei.com/>
- [5] 裁判所, <http://www.courts.go.jp/>
- [6] SynCha: Japanese predicate argument structure analyzer, <http://cl.naist.jp/syncha/>
- [7] Ryu Iida, Kentaro Inui and Yuji Matsumoto, "Anaphora resolution by antecedent identification followed by anaphoricity determination," ACM Transactions on Asian Language Information Processing (TALIP). Vol 4, Issue 4, pp.417-434, 2005.
- [8] 飯田龍, 乾健太郎, 松本裕治, "文脈の手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定," 情報処理学会論文誌, Vol 45, No. 3, 2004.
- [9] 小町守, 飯田龍, 乾健太郎, 松本裕治, "事態性名詞の項構造解析における共起尺度と構文パターンの有効性の分析," 言語処理学会第 13 回年次大会論文集, pp.47-50, March 2007.
- [10] 小町守, 飯田龍, 乾健太郎, 松本裕治, "共起用例と名詞の出現パターンを用いた動作性名詞の項構造解析," 言語処理学会第 12 回年次大会論文集, pp.821-824, March 2006.
- [11] 新里圭司, 黒橋禎夫, "クエリの語句の重要度と係り受けを考慮した自然文検索," 2009 情報処理学会研究報告. No.2, pp.113-120, January.2009.
- [12] 黒橋禎夫, 日笠亘, 藤井綱貴, "入力質問と知識ベースとの柔軟なマッチングに基づく対話的ヘルプシステム," 2001 情報学シンポジウム論文集.
- [13] 原田実, 鈴木亮, "意味グラフのマッチングによる事故問い合わせ文からの判例検索システム JCare," 2001 情報処理学会自然言語処理学会研究報告.No.01- NL-142, pp.15-22, March.2001
- [14] 原田実, 水野高宏, "EDR を用いた日本語意味解析システム SAGE," 2001 人工知能学会論文誌, Vol16, No.1, pp.85-93, January.2001.
- [15] 田淵和幸, 原田実, "日本語文書の意味解析システム SAGE の高速化と精度評価," 2002 言語処理学会第 7 回年次大会論文集, Vol.43, No.9, pp.2894-2902, September.2002.
- [16] 川端崇央, 原田実, "日本語文間の意味的解析システム InSeRa の開発研究," 2001 情報処理学会自然言語処理研究論文集, No.20, pp.105-112, March.2001.
- [17] 江越裕紀, 片上大輔, 新田克己, "判例の構造を利用した判例文書検索(セッション 1:構造化文書による検索応用)," 2005 情報処理学会研究報告. No.11, pp1-8, January.2005.
- [18] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aaron Gull, and Marianna Lau, "Okapi at TREC," 1992 In Text Retrieval Conference, pp. 21-30, 1992.
- [19] LEX/DB インターネット TKC 法律情報データベース, <http://www.tkclex.ne.jp/>