

# 論文閲覧を支援する試み — 文脈検索可能な NLP 予稿集コーパス構築

NaCSE <http://www.mintap.com/nacse/nacse.html>



ミント音声教育研究所 田淵 龍二

[tabuchiryuji@nifty.ne.jp](mailto:tabuchiryuji@nifty.ne.jp)



言語処理学会第 24 回年年次大会(NLP2018)  
2018 年 3 月 14 日 (水) 10 : 30 ~

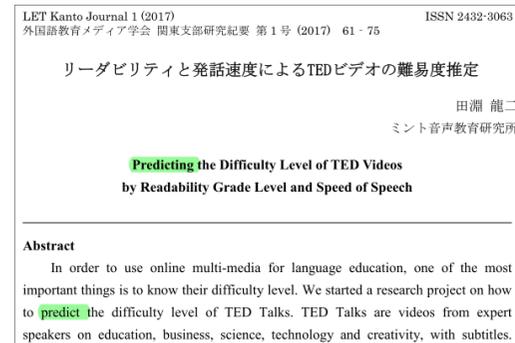


# predict か estimate か

## 査読意見

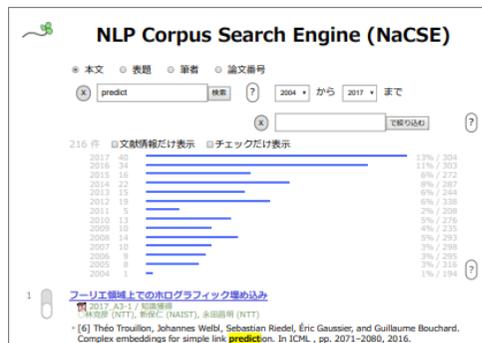
The difficulty level を「見積もる」の意で動詞 estimate を用いているが、間違いでないにせよ、一般的には predict 「予測する」を用いるべきではないか。

## 修正論文



[http://mintap.kir.jp/public/news/pic/let\\_k1\\_2\\_201703.pdf](http://mintap.kir.jp/public/news/pic/let_k1_2_201703.pdf)

## predict 216 hits



## estimate 285 hits



estimating を考慮し estimat で検索

# 発表の流れ

## 1. 研究の動機と目的

自然言語学会（NLP）の研究動向や用語が知りたい  
しかし・・・

グーグルスカラーや J-STAGE は見通しが悪い

## 2. 予稿集専用の検索サイトを構築しよう

仕組み

## 3. 検索例

3-1. アノテーション

3-2. コーパスと著作権

## 4. 長所と欠点

## 5. 著作権について

## 2. 予稿集コーパスの仕組み (1)

### 設計理念

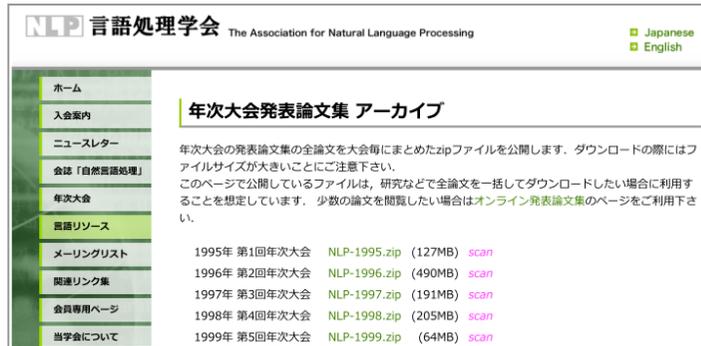
- A. 望む情報を記載した論文か否かを素早くスキャン
- B. 文脈検索可能で一覧性に富む
- C. 原本へのリンクを張るオープンな検索サイト

### 開発手順

1. NLP サイトのリンクから発表論文収集
2. 収集した pdf ファイルからテキスト txt 抽出
3. NLP サイトの年度ごと論文一覧から, 年度・表題・  
筆者情報を収集
4. 上記情報を統合して検索サイト構築

# 2. 予稿集コーパスの仕組み (2) 開発手順

## 1 発表論文 pdf 収集



[http://www.anlp.jp/resource/annual\\_meeting.html](http://www.anlp.jp/resource/annual_meeting.html)

## 2 テキスト txt 抽出



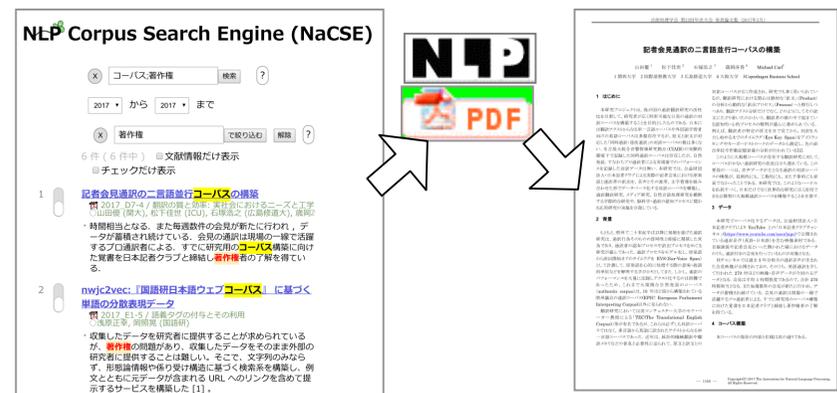
wondershare

## 3 年度 表題 筆者情報収集



[http://www.anlp.jp/proceedings/annual\\_meeting/2017/index.html](http://www.anlp.jp/proceedings/annual_meeting/2017/index.html)

## 4 検索サイト構築



<http://www.mintap.com/nacse/nacse.html>

# 2. 予稿集コーパスの仕組み (3) 文脈検索手法・使い方

## 1 検索語彙入力

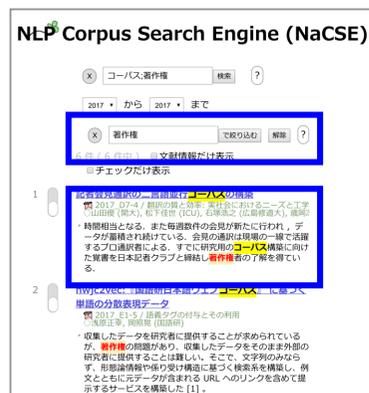


<http://www.mintap.com/nacse/nacse.html>

## 2 段落ごとと彩色出力



## 3 絞り込んでリンクから 4 pdf 文書を開く



[http://www.anlp.jp/proceedings/annual\\_meeting/2017/pdf\\_dir/D7-4.pdf](http://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/D7-4.pdf)

# 3-1. 検索例 アノテーションとタグの違い

検索語： アノテ

絞り込み： タグ

結果： 95 件（423 件中）

- 検索語を「アノテ」としたのは、「～ション」「～ト」などにヒットさせるため
- 「アノテーション」は増加傾向であると判明

言語処理学会 年次大会発表論文集(予稿集): The Association for Natural Language Processing  
 ミント音声教育研究所: Mint Phonetics Education Institute

**NLP Corpus Search Engine (NaCSE)**

検索: アノテ (2004 から 2017 まで) タグ

95 件 (423 件中) 文献情報だけ表示 チェックだけ表示

年	件数
2017	72
2016	60
2015	60
2014	45
2013	38
2012	34
2011	21
2010	26
2009	17
2008	17
2007	13
2006	4
2005	9
2004	7

2 医療テキスト解析のための事象性判定と融合した病名表現認識器  
 2017 A2-1 / 国語表現解析  
 矢野高, 若森翔子, 荒井浩由 (NAIST)  
 \*ここで、<N> タグの精度は、  
 タグのそれと比べて低く、P/N 分類がむしる困難な課題であることを示唆している。P/N 分類は時に、アノテーションの際に

4 日本語 wikification ツールキット: jawikify  
 2017 A2-3 / 固有表現解析  
 松田研史, 岡崎直哉, 乾健太郎 (東北大)  
 \*関係の拡張固有表現に定義された固有表現クラスがアノテされた拡張固有表現タグ付きコーパス [8] (以下、ENE コーパスと表記する) をベースとして、その一部 (およそ 340 記事の新聞記事) に対してエンティティ情報が付与されたコーパス (日本語 wikification コーパス)

8 文の潜在グラフ表現の学習: ニュール機械翻訳での検証  
 2017 A7-4 / 機械翻訳  
 橋本和真, 前田隆雄 (東大)  
 \*れている。このタグ付け器は、人手のアノテーションに基づいて学習することも可能であり、次の係り受け解析層に入力することで学習することも可能である。

20 中国語名詞句の内部構造について  
 2017 C1-3 / 言語資源・コーパス  
 岡坂 (東北大), Alastair Butler (国語研), 吉本啓 (東北大/国語研)  
 \*筆者たちは、中国語の無制約的テキストに対して、統語構造 (ツリー) と論理意味表示 (述語論理式) を付加した中国語の統語・意味表示コーパスを構築している (Butler 2015 など) が、名詞句をめぐるアノテーション作業は二つの課題に直面している。すなわち、(1) 名詞句の内部構造を解明しその解析を決定すること、および (2) 孤立語である中国語には明示的な名詞の格の変化がないため、様々な現れる名詞句の文法的役割を検討し適切な機能タグを付与することである。(2) を徹底的に論じるためには名詞句を中国語の各構文での用法に即して考察する必要があると思われるが、一つの論文でその全部を論じることは無理がある。そこで、本論文はまず (1) を課題とし、名詞句の内部構造を明らかにした上でその解析を決定しようとする。  
 \*単列用 "の主要部もそれぞれ自己"と"列用"になるべきだと考えられる。しかし、実際のところ、中国語の場合、修飾部が主要部名詞の後ろに来ることもあり得る (その詳細は、周他 (2016) を参照のこと)。しかも、本研究のアノテーション方式から見ても、同格関係を内蔵する名詞句の修飾部が前置でなく後置しているという考え方を取るべき理由がある。従来の統語論研究は、"自己"のような再帰代名詞を特別なものとして取り扱うことが多い。そのため、我自己"と"冠単列用"では、主要部名詞と修飾部名詞とがみな同格関係を持ったとしても、両者の区別がツリーバンクの内部でははっきりできるようにしたい。実際のところ、ペン通時コーパス式の解析規約においても、両者の区別がなされている。このような区別は、常に明示的な機能タグの付与によって実現されている。とはいうものの、機能タグが持てるのは句レベルのカテゴリーのみなので、修飾部名詞の名詞句への投射が求められるようになる。さらに、本研究のアノテーション方式では、単一語修飾部の句への投射が認められるのは、それがヘッダの後ろに現れる場合に限られる。従って、"我自己"と"冠単列用"に対しては、修飾部を担当する名詞をそれぞれ二番目の"自己"と"列用"に設定したほうが適当だとと思われる。以上の議論を踏まえて、本研究では、両者の統語構造を次の (9), (10) のように与える。

## 3-2. 検索例 アノテーションとは

検索語： アノテーションとは

絞り込み： —

結果： 3件(6件中3件チェック)

- 「一般にテキストデータに対するアノテーションとは」などの解説を得た
- 「アノテーション」は作業（の結果）、「タグ」は書式だと推察された

言語処理学会 年次大会発表論文集(予稿集): The Association for Natural Language Processing  
 ミント音声教育研究所: Mint Phonetics Education Institute

### NLP Corpus Search Engine (NaCSE)

検索  から  まで  
 で絞り込む

3件 (6件中3件チェック)  文献情報だけ表示  チェックだけ表示

年	件数
2017	2
2016	1
2015	1
2014	1
2013	1
2012	1
2011	1
2010	1
2009	1
2008	1
2007	1
2006	1
2005	1
2004	1

4  [単語単位の日本語係り受け解析 \(pp.955-958\)](#)  
2012\_C4-4 / 構文解析  
 ©Finnerty, Daniel (京大), 高橋拓介 (NII), Neubig, Graham, 森徳介 (京大)  
 ・単語係り受けにおけるフルアノテーションとは、文中の全ての単語に対し係り先を付与することである。一方、部分的アノテーションは、文中の一部の単語のみ係り先を付与することである。この方法は精度向上に貢献しないと推測される係り受けや、種々の持たない係り受けをアノテーションしないため、より効率的であると考えられる。各アノテーション法の例を図 1 に示す。

5  [日本語フレームネットの全文テキストアノテーション: BCCWJへの意味フレーム名付の試み \(pp.703-704\)](#)  
2011\_P3-32 / ポスター  
 ©小原宗子 (慶大)  
 ・日本語フレームネットでは語彙項目アノテーションと全文テキストアノテーションという二つのモードで BCCWJ へのタグ付けを行ってきた。語彙項目アノテーションとは、語彙項目ごとに BCCWJ の中からアノテーション  
 ・対象とする例文をタグ付けしていくモードである。これに対して全文テキストアノテーションとは、特定のサンプルテキスト内の全ての文の、意味フレーム（言語の発話や理解の原に必要となる、体系的知識構造）を喚起（evoke）する全ての語彙項目に対してタグ付けしていくモードを指す。これまで語彙アノテーションでは BCCWJ モニター公開データ 2008 年度版を、全文テキストアノテーションでは BCCWJ コアデータ（人手で形態素解析結果）

6  [アノテーションガイドラインの管理を行う半自動的アノテーションシステムの提案 \(pp.536-539\)](#)  
2008\_D3-6 / テキストDB・言語資源  
 ©大内田賢太, Jin-Dong Kim, 辻井朋一 (東大)  
 ・近年、計算言語学の世界では、大量のテキストデータ（コーパス）が蓄積されるようになってきたことから、それらのコーパスに対して様々な情報を付与（アノテーション）し、アノテーションされたコーパスから言語処理用知識を得る手法が一般的に用いられている。それゆえ、コーパスのアノテーションは計算言語学の世界で重要なテーマの1つになっている。一般にテキストデータに対するアノテーションとは、テキストデータ中の単語もしくは単語列を指定し、指定された単語・単語列に何らかの情報を付与することである。人手によるアノテーションにおける問題の一つとして、アノテーションの一貫性の維持の困難さがあげられる。ある情報をテキストに付加するとき、同じ情報をアノテーションするときでも、単語・単語列の領域の指定の仕方にずれが生じたり、どの単語・単語列にアノテーションしたらいいのか判断が難しい場合がある。加えて、アノテーション作業は非常に多くの時間を要し、しばしば数週間、数か月かかる。そのため、異なるアノテーター間で一貫性が損なわれる危険性が常につきまってくる（inter-annotator discrepancy）。それどころか、1人1人のアノテーションを行う人同士（アノテーター）においても、時間の経過につれて一貫性が狂う可能性がある（intra-annotator discrepancy）。アノテーターはどのようにアノテーションしたらいいのか悩んだ場合、一貫性を維持するためにアノテーションガイドラインを参照することになる。アノテーションガイドラインとは、アノテーター同士で決めておくアノテーション方針である。そのため、アノテーションガイドラインの作成が重要な課題になる。

# 3-3. 検索例 コーパスの著作権処理

検索語：著作権;コーパス

↑ 複合検索 (and / or)

絞り込み： —

結果： 59 件

- コーパスの著作権問題が議論されていることがわかった
- 研究ごとに著作権処理方法の特徴があった
- 後ほど 5 で詳しく触れる

言語処理学会 年次大会発表論文集(予稿集): The Association for Natural Language Processing  
 ミント音声教育研究所: Mint Phonetics Education Institute

### NLP Corpus Search Engine (NaCSE)

◎ 本文 ◎ 表題 ◎ 筆者 ◎ 論文番号

◎ コーパス:著作権 検索 ? 2004 から 2017 まで

◎ で絞り込む ?

59 件 ◎ 文献情報だけ表示 □ チェックだけ表示

2017	6	2% / 304
2016	11	4% / 303
2015	8	3% / 272
2014	6	2% / 287
2013	8	3% / 244
2012	2	1% / 338
2011	3	1% / 208
2010	3	1% / 276
2009	3	1% / 235
2008	3	1% / 293
2007	3	1% / 298
2006	2	1% / 295
2005	0	0% / 316
2004	1	1% / 194

- 1 **記者会見通訳の二言語並行コーパスの構築**  
 ◎ 2017\_D7-4 / 翻訳の質と効率: 実社会におけるニーズと工学的実現可能性  
 ◎ 山田俊 (阪大), 松下佳世 (ICU), 石塚清之 (広島修道大), 渡岡啓吾 (阪大), Carl Michael (CBS)
- 2 **nwjc2vec: 『国際研日本語ウェブコーパス』に基づく単語の分散表現データ**  
 ◎ 2017\_E1-5 / 種別タグの付与とその利用  
 ◎ 渡部正幸, 岡崎寛 (国語研)
- 3 **パターンに基づく統計翻訳とその未知語処理**  
 ◎ 2017\_P11-7 / ポスター  
 ◎ 川原等, 村上仁一 (鳥取大)  
 \* 本実験には、電子辞書などの例文より抽出した単文コーパス [3] を用いる。使用するデータの内訳を表 1 に示す。  
 \* 5.2 コーパスの前処理  
 \* 本実験では、単文コーパスに対して前処理を行う。具体的には、日本語文に対して“MeCab”による形態素解析を行い、英語文に対して“tokenizer.perl”による分詞処理を行う。前処理を行った単文コーパスの例を表 2 に示す。  
 \* 表 2 前処理を行った単文コーパスの例  
 \* [3] 村上仁一, 藤波進 “日本語と英語の対訳文対の収集と著作権の考察”, 第一回コーパス日本語学ワークショップ, pp.119-130. 2012.
- 4 **人手対訳包を利用したパターンベース統計翻訳**  
 ◎ 2017\_P15-8 / ポスター  
 ◎ 安場裕人, 村上仁一 (鳥取大)  
 \* 5.1.1 単文コーパス  
 \* 実験には電子辞書などの例文より抽出した単文コーパス [5] を用いる。使用するデータの内訳を表 1 に示す。  
 \* [5] 村上仁一, 藤波進 “日本語と英語の対訳文対の収集と著作権の考察”, 第一回コーパス日本語学ワークショップ, pp.119-130. 2012.
- 5 **ゲーミフィケーションを利用した対話ログ収集における応答文の改善と対話ログの解析**  
 ◎ 2017\_P16-1 / ポスター  
 ◎ 村内康, 尾形順哉, 金子正弘, 河村純策, 北川高彬, 黒田祐司, 奥藤宏行, 山本豊, 小町守 (前橋大)  
 \* いる。非タスク指向型対話システムにおいては、Twitter を利用することで、日本語における (1) 大規模で (2) において Tweet と Reply の関係を大量に収集してきて公開可能な (3) ラベル付きの雑談対話コーパスの開発を対話コーパスとして利用する研究 [5, 8] や大規模な映画目標した。ゲーミフィケーションによってユーザにチャット  
 \* の字幕データを対話コーパスとして利用した研究 [1, 3] がある。しかし、英語では対話コーパスが充実しているものの、日本語における共通して利用可能な大規模  
 \* トボットの育成をしてみたら副産物として、雑談対話コーパスを作成するシステムを提案している。ゲームにおいて、各ユーザは与えられた発話文に対する応答文を選択

# 3-4. 使い方手順

1. 検索対象を選ぶ
2. 検索年の範囲を選ぶ
3. 検索語を記入する
4. 検索ボタンをクリック



## 配置

表題  
文献 pdf へのリンク

チェックボタン  
気になった文献を  
マークしておく

**NLP Corpus Search Engine (NaCSE)**

● 本文 ● 表題 ● 筆者 ● 論文番号

✕ コーパス,著作権 検索 ? 2013 ▼ から 2015 ▼ まで

✕ で絞り込む ?

22 件 (1 件チェック) ■ 文献情報だけ表示 ■ チェックだけ表示

2015	8	3%	272
2014	6	2%	287
2013	8	3%	244

1 [ロジスティック回帰分析を用いたパターンに基づく統計翻訳\(pp. 245-248\)](#)

2015\_A1-1 / 機械翻訳  
○ 春野瑞季, 村上仁一, 徳久雅人 (鳥取大)

2 [RBMT, SMT, Hybrid MTの特徴比較と今後の展望 \(pp. 249-252\)](#)

2015\_A1-2 / 機械翻訳  
○ 高宗徹, 真下修三, 趙東社, 川上健 (高電社)

? ヘルプ

表題  
論文番号  
筆者

論文情報  
ヒット文脈

操作エリア  
統計エリア  
結果エリア

## 4. 長所と欠点 (1) 欠点

仕組みに由来する限界

### 1. pdf からテキスト抽出時に文字化け > 検索失敗

文節抽出型文簡約における読みやすさ向上のための文節末修正



節□□□□約□□□□□□□□向□□□□□節末修□

2006\_C5-5.pdf

### 2. 段落抽出時に語彙が改行 > 検索失敗

力し始めるまでのタイムラグ ( Eye-Key Span ) をアイトラック ↩ 改行 [hard return](#)

キングやキーボードストロークのデータから測定し、先の訳

2017\_D7-4.pdf

### 3. 日本語仕様なので、見出し語検索には未対応

## 4. 長所と欠点 (2) 対策

1. pdf からテキスト抽出時に文字化け > 検索失敗

**対策** 表題で検索

2. 段落抽出時に語彙が改行 > 検索失敗

**対策** 長いカタカナ語は短くして検索

例えば「アノテーション」であれば「アノテ」

3. 日本語仕様なので、英語の見出し語検索には未対応

**対策** コアになる文字列で検索

例えば「estimate」であれば「estim」

## 4. 長所と欠点 (3) 長所

### 実現した設計理念

- A. 望む情報を記載した論文か否かを素早く判断
- B. 文脈検索可能で一覧性と視認性に富む
- C. 原本へのリンクを張るオープンな検索サイト

### 長所と今後の課題・展望

1. 資源を予稿集としたので論文より情報が豊富で早い
2. 仕組みが簡明なので他の予稿集にも手軽に対応可能
3. リンクによる検索エンジン仕様なので著作権クリア
4. 【課題】 検索用テキストの正確性向上

## 5-1. 著作権処理方法

### 言語処理学会予稿集に見る著作権処理方法の類型

- type 1: 著作権処理を実施  
時間と人件費がかかる
- type 2: 著作権の切れた著作物  
映画や青空文庫
- type 3: コーパス未公開
- type 4: n-gram データを頒布
- type 5: 元著作物へのリンクを公開  
対象著作物がウェブ公開の場合
- type 6: 疑似テキストを自動生成して公開
- type 7: 公開可能な著作物

## 5-2. 著作権処理方法 予稿集からの引用

### type 1: 著作権処理を実施

代表性を有するコーパスの設計とサンプリングの実際 —コーパスに基づく言語研究の可能性と限界— (pp. 150-153) / 2010\_S1-3 / 「言語表現」と「言語」のあいだ / 丸山岳彦 (国語研) / すべてのサンプルについて著作権処理を実施し、研究者間で共有されるコーパスとしての公開を目指している。

### type 2: 著作権の切れた著作物

「青空文庫」を言語コーパスとして使おう—メタデータ構築による歴史的・社会言語学的研究への応用の試み— (pp. 915-918) / 2006\_P7-11 / ポスター / 千葉庄寿ほか / 死後 50 年で著作権の切れる 1956 年 (2006 年現在) までに死去した作家 300 人あまりの作品が収録されている

### type 3: コーパス未公開

意味的逆引き辞書『真言』 (pp. 406-409) / 2013\_P2-2 / ポスター / 栗飯原俊介(九大)ほか / 前提とする辞書データに著作権があるため、現在の実装は未公開 / 「CD-毎日新聞データ集」に含まれるデータの特徴について (pp. 878-881) / 2013\_P6-16 / ポスター / 〇長谷川守寿(首都大) / 現在著作権交渉中の為、本文は表示できません

### type 4: n-gram データを頒布

『国語研日本語ウェブコーパス』の検索系 / 2016\_P17-7 / ポスター / 浅原正幸 (国語研) ほか / Web コーパスの先行事例では、著作権などの問題を回避するために n-gram データを頒布しているものが多い。

### type 5: 元著作物へのリンクを公開

nwjc2vec: 『国語研日本語ウェブコーパス』に基づく単語の分散表現データ / 2017\_E1-5 / 語義タグの付与とその利用 / 浅原正幸, 岡照晃 (国語研) / 収集したデータを研究者に提供することが求められているが、著作権の問題があり、収集したデータをそのまま外部の研究者に提供することは難しい。そこで、文字列のみならず、形態論情報や係り受け構造に基づく検索系を構築し、例文とともに元データが含まれる URL へのリンクを含めて提示するサービスを構築した

### type 6: 疑似テキストを自動生成して公開

ゲーミフィケーションを利用した対話ログ収集における応答文の改善と対話ログの解析 / 2017\_P16-1 / ポスター 叶内農ほか / 公開を最終目標にしているため、著作権に配慮し、応答文の生成を全て DoCoMo の雑談対話 API に頼っていた。

クラウドソーシングにおける成果物の品質維持のためのダミー問題 出題手法の検討 (pp. 678-681) / 2014\_C4-1 / 言語資源・コーパス / 清水伸幸ほか / 実際の広告著作権上の問題で実験での利用が困難であるため、広告の代わりに以下の様な擬似的な特集記事の見出しを自動生成し、

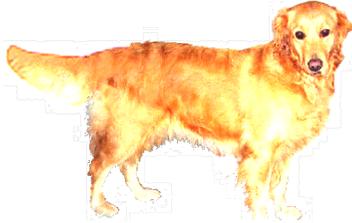
### type 7: 公開可能な著作物

「拡張固有表表現+Wikipedia」データ / 2016\_P2-4 / ポスター / 関根聡ほか /

この辞書の問題は著作権も絡み非常に根深いものではあるが、特に固有表現を対象としたクラウドソーシングで作られた Wikipedia は、この問題を解決する一つの有力なリソースであると言える。

意味検索のプロトタイプシステムの構築 (pp. 823-826) / 2012\_P2-24 / ポスター / 大倉清司ほか / 著作権やデータの入手の容易性から本研究では特許明細書のテキスト部分を対象

ありがとうございました



アンケートは回収箱へ

ハンドアウト（PDF 版）や TED コーパスな  
どの記事を配信するメルマガをご希望の方は  
メールでお知らせください

[tabuchiryuji@nifty.ne.jp](mailto:tabuchiryuji@nifty.ne.jp)