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

Chrome
(recommended)

ミント音声教育研究所　田淵　龍二
tabuchiryuji@nifty.ne.jp

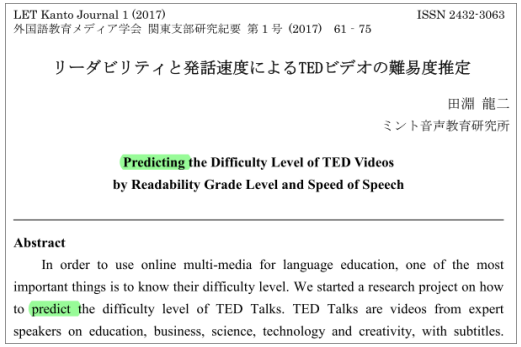
言語処理学会第 24 回年年次大会(NLP2018)
2018 年 3 月 14 日（水）10：30～

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



LET Kanto Journal 1 (2017) ISSN 2432-3063
外国語教育メディア学会 関東支部研究紀要 第 1 号 (2017) 61 - 75

リーダビリティと発話速度によるTEDビデオの難易度推定

田淵 龍二
ミント音声教育研究所

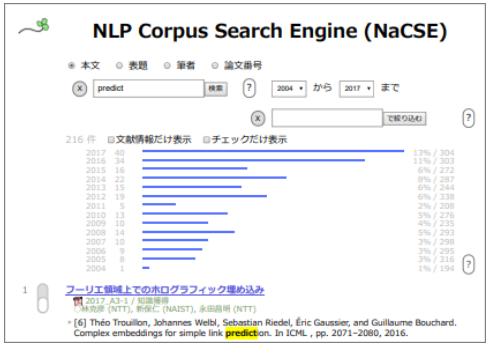
Predicting the Difficulty Level of TED Videos by Readability Grade Level and Speed of Speech

Abstract

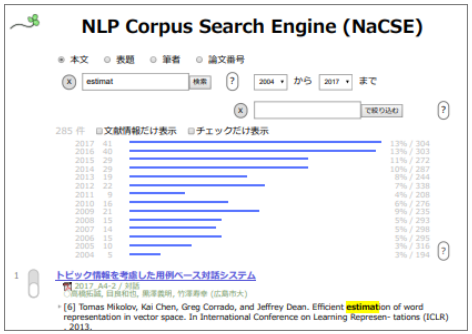
In order to use online multi-media for language education, one of the most important things is to know their difficulty level. We started a research project on how to predict the difficulty level of TED Talks. TED Talks are videos from expert speakers on education, business, science, technology and creativity, with subtitles.

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

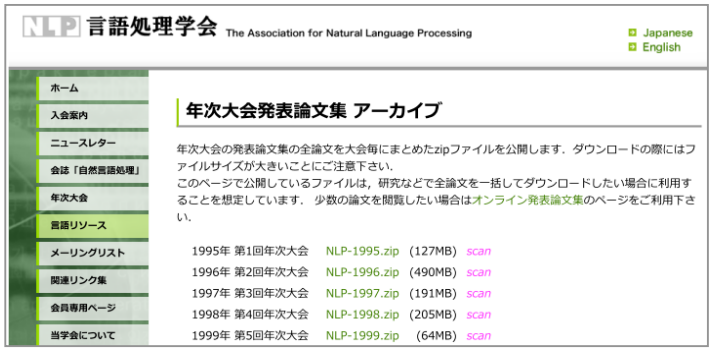
C. 原本へのリンクを張るオープンな検索サイト

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

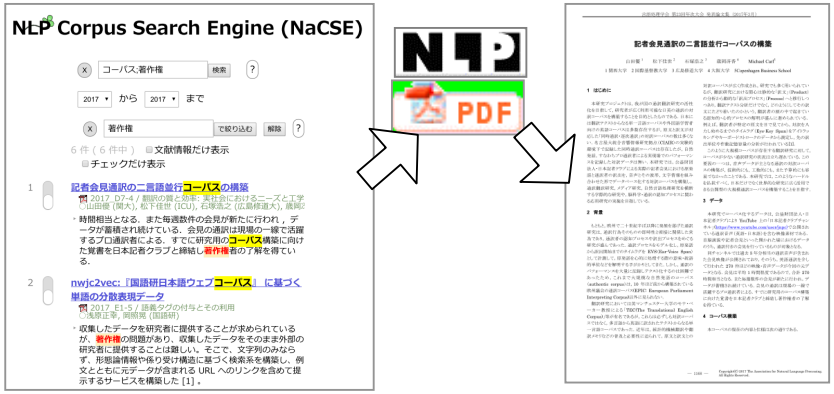
2 テキスト txt 抽出

wondershare

3 年度 表題 筆者情報収集

http://www.anlp.jp/proceedings/annual_meeting/2017/index.html

4 検索サイト構築



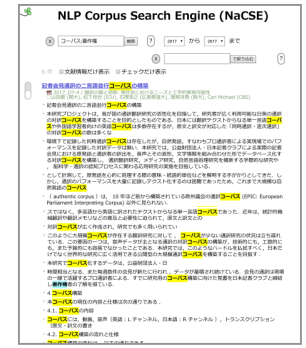
http://www.mintap.com/nacse/nacse.html

2. 予稿集コーパスの仕組み (3) 文脈検索手法・使い方

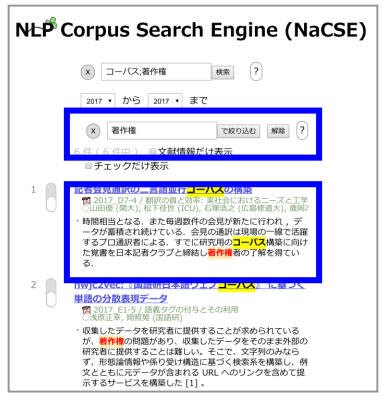
1 検索語彙入力

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

3-1. 検索例 アノテーションとタグの違い

検索語： アノテー
絞り込み： タグ
結果： 95 件（423 件中）

- 検索語を「アノテー」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

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

3-3. 検索例 コーパスの著作権処理

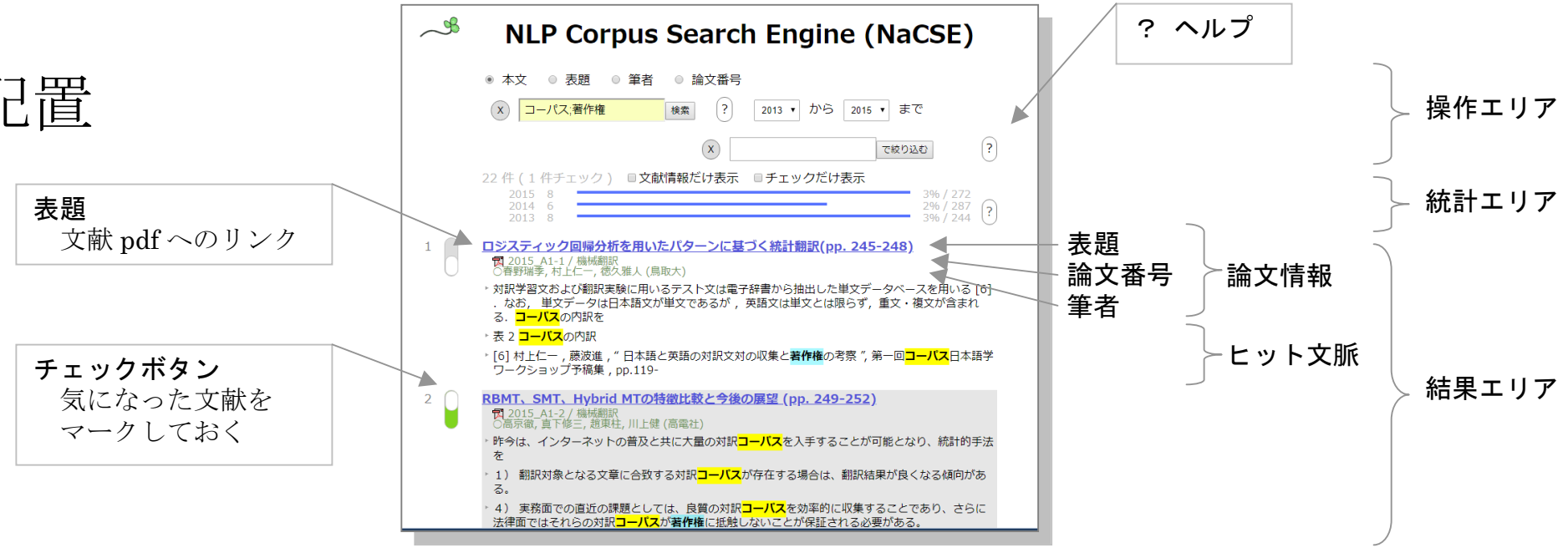
検索語：著作権;コーパス
↑複合検索（and / or）
絞り込み： ー
結果： 59 件

- 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

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



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⏎ 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 / 粟飯原俊介(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 / ○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 叶内晨ほか /
公開を最終目標においているため, 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 /
実際の広告は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