学術論文検索エンジンの要件のひとつ ― 文脈検索結果

田淵　龍二　（ミント音声教育研究所）

はじめに

コーパス構築と検索エンジン開発において，先行研究者たちが言語資源の著作権処理をどのように解決しているかを知るために，各種検索サイトを利用して調査した。しかし既存のサイトでは満足な生産性が得られなかった。そこで能率的に作業できる新しい検索エンジンを開発した。この論文はその顛末記でもある。

先行研究と目的

ウェブに公開された膨大な論文資源から必要情報を手早く取得する仕組みとしては，論文間の参照情報を利用したサーベイ論文作成支援システム（難波・奥村, 1999），サーベイ論文自動作成（飯沼・難波・竹澤, 2014）や，自動知識抽出システム（澤山・鈴木・進藤・松本, 2018），類似論文検索のための論文ベクトル学習法（小林・松本, 2018）などが考案されている。今回は，ウェブに公開された論文を使って，初学者でも手早く効果的に学術的調べ物ができる検索エンジン（検索サイト）の要件について研究した。

方法

日本で一般的な CiNii，JAIRO，J-STAGE，Google Scholar の 4 つのサイトについて，所望論文到達性能を調べた。性能検査では仮の探求項目を「対訳コーパス検索エンジン」とし，「対訳コーパス」と「エンジン」を検索語彙とした。

結果

CiNii ではヒット論文の表題と抄録引用などの一覧が表示され，ひとつを選択すると詳細な書誌情報のページが開く。しかし，抄録だけでは「対訳コーパス」と「エンジン」のどのような関連について述べているのか，あるいはたまたま語彙があっただけなのかは不明であった。所望論文か否かの判定には論文閲覧が必要となる。そこで論文を探すが，ウェブで公開されていないことも多い。

JAIRO では該当項目なしであった。

J-STAGE は CiNii とほぼ同等であった。

Google Scholar では表題と書誌情報の 2 行の下に 3 行分の引用があり，検索語が太字で強調表示されていたので，検索語を含む文脈がひと目でわかり，すばやく取捨選択できた。しかし引用文は 3 行しかなく，文意理解の正確さに欠けた。

考察

以上より，限りある時間内で効果的に学術的な調べ物をする研究者，特に初学者を支援するサイトとしては，以下の要件が重要であると判明した。

1. 視認性： 検索語が強調表示されること
2. 文脈性： 検索語を含んだ文脈引用が十分にあること
3. 一覧性： 検索結果のヒット項目が一望できること

4. 絞込み： 検索結果項目をさらに絞り込めること
5. 選択性： 注目した論文をマークできること
6. 閲覧性： ワンステップでリンク先論文を閲覧できること

課題解決

これらの要件に合致した論文検索エンジンが見当たらなかったので，新しく NaCSE （Natural Language Processing Corpus Search Engine，ナックス）を構築した。これは長年に渡ってコーパス構築と検索エンジン作成を行ってきた言語処理学会の学会誌と年次大会発表論文集あわせて 24 年分 4,428 本を対象としたものである。

主検索語をコーパス，副検索語を著作権とした場合の出力の様子を図 1 に示す。最上部に検索語の入力欄がある。その下の青い横棒は年次ごとのヒット論文数で，その下から結果項目が年次順に並ぶ。主検索語は黄色，副検索語は橙色で強調されている。ヒット論文数は 2,322 本（5 割）で，うち 50 本のヒット文脈で著作権の併記があった。

図 1. NaCSE での検索結果表示の様子
主検索語はコーパス，副検索語は著作権

まとめ

50 本の論文を解析したところ，著作権処理はコーパスごとに特徴があり，7 種類に整理できた。それによると筆者が開発した映画セリフ検索サイト Seleaf は「著作権の切れた著作物」，TED コーパス selected360 と論文コーパス NaCSE は「引用元へのリンク」，文法コーパス SCoRE on Talkies は「許諾済」に該当した。実証実験では，検索から所望論文閲覧まで円滑に進み，高い生産性が実証された。

参考文献・参考サイト

飯沼俊平・難波英嗣・竹澤寿幸 (2014).「新情報の追加によるサーベイ論文の作成支援」言語処理学会第 24 回年次大会発表論文集, 408–411.

小林雄太・松本裕治 (2018).「論文の構成要素を考慮した分散表現に基づく類似論文検索」言語処理学会第 24 回年次大会発表論文集, 959–962.

NaCSE (2018). http://www.mintap.com/nacse/nacse.html

難波英嗣・奥村学 (2018).「論文間の参照情報を考慮したサーベイ論文作成支援システムの開発」自然言語処理, *6*, 43–62.

澤山熱気・鈴木潤・進藤裕之・松本裕治 (2018).「非即時的なタスク設定における固有表現抽出の改善」言語処理学会第 24 回年次大会発表論文集, 921–924.

田淵龍二 (2018).「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」言語処理学会第 24 回年次大会発表論文集, 686–689.