Vol.2019-DBS-169 No.20
Vol.2019-IFAT-136 No.20
2019/9/10

NTCIR-15 ウェブ検索・再現可能性タスク (WWW-3) および対話評価タスク (DialEval-1) への誘い

酒井 哲也[1,a)]

概要：本稿の目的は，NTCIR-15 に向けて採択された We Want Web with CENTRE (WWW-3) タスクおよび Dialogue Evaluation (DialEval-1) タスクへの参加を読者に検討していただくことである．WWW-3 は古典的なウェブ検索タスクであり，公開されている Anserini などの検索ツールキットなどを用いて容易に参加できる．また，CENTRE というサブタスクでは，他の研究者が発表した実験結果を追試する形での参加が可能である．一方，DialEval-1 は NTCIR-14 Short Text Conversation タスクを踏襲したヘルプデスク対話の品質を予測するタスクであり，多様な顧客の相手をする対話システムの自己診断技術の深耕を狙ったものである．各タスクについて，これまでの取り組みの概要と，NTCIR-15 タスク参加者に求められることを説明する．

Invitation to the NTCIR-15 We Want Web with CENTRE (WWW-3) and Dialogue Evaluation (DialEval-1) Tasks

Tetsuya Sakai[1,a)]

Abstract: The objective of this paper is to try to convince the reader to participate in the We Want Web with CENTRE (WWW-3) Task and the Dialogue Evaluation Task, which have been accepted for NTCIR-15. WWW-3 is a classical web search task, and it is easy to participate in it by utisiling a publicly available information retrieval toolkit such as Anserini. In CENTRE, a subtask of WWW-3, participants can try to reproduce the results previously reported by other researchers. On the other hand, DialEval-1 is a continuation of the NTCIR-14 Short Text Conversation Task and deals with predicting the quality of a given helpdesk dialogue. The aim of this task is to enable self-diagnosis of dialogue systems that are required to face diverse customers. For each task, we provide a quick summary of the past rounds and a description of what is expected of an NTCIR-15 participant.

1. はじめに

情報アクセスの評価型国際会議 NTCIR (NII Testbeds and Community for Information access Research)[*1] は 1999 年以来約 1 年半毎に開催されており [21]，2020 年 12 月には第 15 回 (NTCIR-15) を迎える．その約 20 年間における研究の新規性に焦点を当てた本が同年に Springer 社より出版予定であり，ドラフト版は既にオンラインで閲読可能である[*2]．

本稿では，NTCIR-15 のタスクとして 2019 年 7 月に採択が決まった We Want Web with CENTRE (WWW-3) タスク[*3] および Dialogue Evaluation (DialEval-1) タスク[*4] を，両タスクのオーガナイザという立場から紹介する．それぞれのタスクに関するこれまでの取り組みと，NTCIR-15 におけるタスク設計を日本語で簡潔に説明し，読者 (特に学生) の方々にタスクへの参加を検討していただくことを目的としている．両タスクとも中国語および英語のデータを扱っているが，日本の研究者が比較的扱いやすいと思われる英語データを使ったサブタスクを中心に説明する．

[1] 早稲田大学 (Waseda University)
[a)] tetsuyasakai@acm.org

[*1] http://research.nii.ac.jp/ntcir/index-ja.html
[*2] http://sakailab.com/ntcirbookdraft/
[*3] http://www.thuir.cn/ntcirwww3/
[*4] http://sakailab.com/ntcir15dialeval1

2. We Want Web with CENTRE (WWW-3)

2.1 前回までのあらすじ: WWW-1, WWW-2

2.1.1 WWW タスク概観

ウェブ検索を明示的に銘打ったタスクとしては，米 TREC (Text Retrieval Conference)[*5]で 1999〜2004 年および 2009〜2014 年に開催された Web Track および NTCIR で 2002〜2005 年 (NTCIR 3〜5) に開催された WEB タスクがある [21]．TREC 2014 年において Web Track が終了となった際，一部の研究者たちがふざけて We Want Web (「ウェブトラックを廃止するな!」) というプラカードを掲げたことをヒントに[*6]，筆者らは NTCIR-13 において We Want Web (WWW) タスクを開始した．

日頃，ウェブ検索エンジンにあまり不満を感じないかもしれないが，現在の検索エンジンはあくまでキーワードマッチングを基本としており，ユーザの細かい情報要求に必ずしも答えられていない．また，ユーザには検索エンジンが返してくるウェブページしか見えないが，検索エンジンが決して返すことのない適合ページが多く存在するかも知れない．すなわちウェブ検索は「解決済の研究課題」ではない．情報検索の研究者であるならば，どのような情報要求に対しどのような手法でどの程度の検索有効性が実現でき，課題は何であるかを明らかにし，さらに，技術進歩を長期的に検証すべきである．We Want Web (WWW) はこのような問題意識のもとに運営されている．

WWW タスクは，初期 TREC においてアドホック検索 [21] と呼ばれたものと基本的に同じである．参加チームは，与えられた各トピック (検索課題) に対し，ランク付き検索結果を生成し，オーガナイザに提出する．この結果ファイルをラン (run) という．WWW の英語サブタスクでは，検索対象コーパスとして，2012 年にクロールされた約 5 千万件のウェブページからなる `clueweb12-B13` というウェブデータを用いている[*7]．このデータは TREC Web Track などでも用いられている．

2.1.2 NTCIR-13 WWW-1 英語サブタスク公式結果

NTCIR-13 WWW-1 では 100 件のトピックが用意された．英語サブタスクに参加したのは，豪 RMIT 大学，中国の清華大学 (THUIR)，中国人民大学 (RUCIR) のみであったが，トップの RMIT のランは THUIR のランを統計的に有意に上回る検索有効性を達成した [7][*8]．ただし，RMIT の手法は，順次依存モデル (sequential dependency model) という ACM SIGIR 2005 で発表された論文の手法 [9] にクエリ拡張 (query expansion) を加えたものであり [3]，新しい手法というわけではなかった．

2.1.3 NTCIR-14 WWW-2 英語サブタスク公式結果

NTCIR-14 WWW-2 では 80 件のトピックが用意された[*9]．英語サブタスクに参加したのは，MPII (独 Max Planck Institute for Informatics)，前述の THUIR および RUCIR，早稲田大学 (SLWWW)，およびオーガナイザチーム (ORG) の 5 チームであった．このうち THUIR のランのいくつかは，他のチームのいくつかのランを統計的に有意に上回る検索有効性を達成した [8][*10]．ただし，THUIR の有効であったランは，LambdaMART，AdaRank，Coordinate Ascent という既存の Learning to Rank 手法を適用したものであり [20]，新しい手法を用いているわけではなかった[*11]．彼らは，MQ2007 および MQ2008 という Learning to Rank データセット [10] を学習用に，WWW-1 の英語データをバリデーションに用いて上記各手法を適用している．

2.2 前回までのあらすじ: CENTRE

2.2.1 CENTRE タスク概観

情報検索研究における最近の話題として，評価実験の再現可能性にまつわる課題がある．本稿では，ACM (Association for Computing Machinery) の定義[*12]を参考に，以下のように用語を定義する．

反復可能性 (repeatability) 同じチームが同じ評価データ上で同じ結果を出せること．

複製可能性 (replicability) 元の実験を実施したチームとは別のチームが，元の実験と同じ評価データ上で同じ結果を出せること．

再現可能性 (reproducibility) 元の実験を実施したチームとは別のチームが，元の実験とは異なる評価データ上で同じ結果を出せること．

これらの性質が，研究コミュニティ全体としての技術向上に不可欠であることは明らかであろう．

NTCIR-14 において実施された CENTRE (**CLE**F **N**TCIR **T**REC **Re**producibility) タスクとは，上記のうち複製可能性と再現可能性に着目した，NTCIR·TREC·CLEF (欧州の NTCIR や TREC に相当する評価型会議)[*13]の協力体制に基づくものであった．CENTRE はこれまで TREC 2018，CLEF 2018 および 2019，NTCIR-14 において開催されてきた．特に NTCIR 版の CENTRE は，高度なラン (A-run: **A**dvanced run) と比較対象のラン (B-run: **B**aseline

[*5] `http://trec.nist.gov/`
[*6] `https://twitter.com/djoerd/status/536128465276530688`
[*7] `http://lemurproject.org/clueweb12/`
[*8] 評価指標 nDCG および Q-measure [21] で測定した場合．
[*9] このトピック数は，トピック数設計 (topic set size design) という統計的手法 [12], [21] により WWW-1 のデータに基づき決定された．
[*10] 評価指標 nDCG および Q-measure で測定した場合．
[*11] THUIR は，中国語サブタスクにおいては大量の学習データを活用した深層学習ベースの手法を提案している [20]．
[*12] `https://www.acm.org/publications/policies/artifact-review-badging`
[*13] `http://www.clef-initiative.eu/`

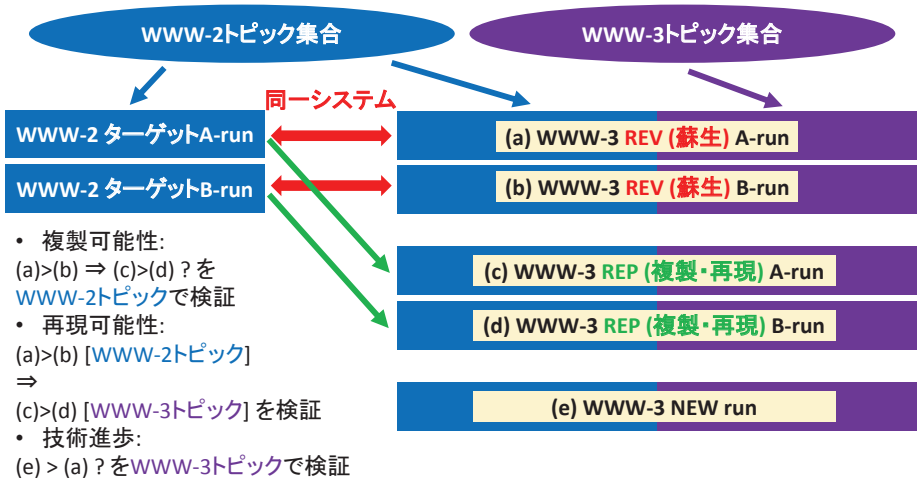

図 1　WWW-3 英語サブタスクの仕組み

run) の間の効果 (effect) [12], [21] を複製もしくは再現することに主眼を置いている．ここでの効果とは，大まかには両ランの検索有効性の差がどれくらいか (標準偏差いくつぶんか) を意味する．

NTCIR-14 CENTRE の複製可能性サブタスクは，NTCIR-13 WWW-1 の RMIT の 2 つのラン [3] の間の効果を，他のチームが (同じデータ上で) 複製できるかを問うものであった．前述の通り，RMIT は WWW-1 において最高成績を収めたチームである．また，再現可能性サブタスクは，TREC 2013 Web Track における米 Delaware 大の 2 つのラン [15] の間の効果を，他のチームが WWW-1 テストコレクション上で再現できるかを問うものであった．後者は，web-based working sets という ACM SIGIR 2006 で提案された検索ターム選定手法 [2] に関するものである．

2.2.2　NTCIR-14 CENTRE タスク公式結果

残念ながら，CENTRE タスクに結果を提出したチームは前述の独 MPII [16] のみであった．他のチームの研究結果の複製もしくは再現のみを行うというタスクは，なかなか研究者の動機付けが難しいのかも知れない．しかし，MPII が提出した結果は，複製可能性，再現可能性のいずれにおいても概ね良好な結果であった．主な結果は以下の通りである [13].

- RMIT による「順次依存モデルがベースラインよりも有効である」という知見およびその効果を，MPII が複製することができた．また，統計的有意差も複製できた．ただし，トピック毎の差においては，RMIT と MPII の間にかなりの隔たりがあった．
- Delaware による「web-based working sets が (TREC データにおいて) ベースラインよりも有効である」という知見およびその効果を，MPII が再現することができた．また，統計的有意差も再現できた．

2.3　WWW-3 のタスク設定

WWW-3 は WWW と CENTRE が統合されたタスクであるが, そのタスク設定は, 参加者の立場からすると WWW-1, WWW-2 と同様である．すなわち，`clueweb12-B13` の検索インデクスを作成しておき，与えられたトピック集合の各トピックについて，ランク付き検索結果を出力すればよい．

図 1 に，WWW-3 英語サブタスクの仕組みを示す．まず，オーガナイザは，複製および再現のための A-run および B-run として，WWW-2 から 2 つのランを指定する．具体的には，A-run として THUIR の LambdaMART に基づくラン，B-run として同チームの BM25 に基づくラン [20] を用いる予定である[*14]．WWW-3 タスクの全参加チームには，図に示すように，WWW-2, WWW-3 トピック集合の両方を一括で与え，各トピックに対する検索結果を作成してもらう．WWW-3 に提出されるランには以下の 3 種類がある．

REV run　Revived (蘇生された) run, すなわち, WWW-2 において A-run, B-run を提出したチームが，再度 WWW-2, WWW-3 トピック集合に対して同じアルゴリズムを用いて検索を行った結果である．

REP run　これは上記以外のチームが A-run, B-run を模して検索結果を作成したものである．このランの WWW-2 トピックに関する部分は複製可能性 (A-run>B-run という結果を同じデータで別のチームが得ることができるか?) の実験結果に相当し，WWW-3 トピックに関する部分は再現可能性 (A-run>B-run という結果を別のデータで別のチームが得ることができるか?) の実験結果に相当する．

NEW run　独自のアルゴリズムにより作成した検索結果．なお，WWW-2 トピックに対する適合性判定 (WWW-2 に

[*14] WWW-2 において両者の間に統計的有意差はなかった．

おいて構築された正解データ) は事前に参加者に与えられる．一方，WWW-3 トピックに対する適合性判定はラン提出時点では存在しない．

図 1 に示したように，(a), (b), (c), (d) の青い部分に着目することにより複製可能性の検証を行うことができ，同様に，(a), (b) の青い部分と，(c), (d) の紫の部分に着目することにより再現可能性の検証を行うことができる．さらに，独自の結果として提出された (e) と，WWW-2 における state-of-the-art である (a) を紫の部分において比較することにより，技術進歩の検証が可能である．既存の検索ツールキット Anserini [6] などを活用し，是非本タスクに挑戦していただきたい．

3. Dialogue Evaluation (DialEval-1)

3.1 前回までのあらすじ: STC-1, STC-2, STC-3

3.1.1 STC タスク概観

もともとの Short Text Conversation は，与えられたツイート (もしくは Weibo[*15]への中国語の投稿) に対し，適切な自然言語のレスポンスを返す，single-round 対話タスクであった．しかし，STC-3 はこれとは大きく異なり，multi-round 対話を扱う以下の 2 つを含む 3 つのサブタスクが扱われた [17][*16]．

対話品質 (Dialogue Quality: DQ) 与えられたヘルプデスク・顧客間の対話に対し，20 名程度の判定者による 5 段階の対話品質評価データを作成しておく．このスコアの分布を正解データとし，参加システムはこの分布を予測する．

ナゲット検出 (Nugget Detection: ND) 同対話中の各ターン (turn) に対し，20 名程度の判定者によりナゲットの種類を分類したデータを作成しておく．ここで，ナゲットとは，問題を抱えている顧客が，問題解決に至るための状態遷移に貢献する情報を含む文字列を意味する [17], [18]．上記のナゲット分類結果の分布を正解データとし，参加システムはこの分布を予測する．

いずれも対象言語は中国語および英語である．英語サブタスクでは，Weibo からマイニングした中国語の対話を人手により英語に翻訳したデータを用いている．

対話の品質としては以下の 3 種類を定義している．

A-score タスク達成度 (task **A**ccomplishment)
S-score 対話顧客満足度 (customer **S**atisfaction)
E-score 対話有効性 (dialogue **E**ffectiveness)

各判定者は，−2 点から 2 点までの五段階評価でスコアを与える．

一方，ND サブタスクにおける各ターンは，各判定者により以下のいずれか 1 つに分類される．

CNUG0 トリガ・ナゲット．顧客が直面している問題をヘルプデスクに伝えているターンを意味する．
HNUG, CNUG それぞれヘルプデスク・顧客の通常のナゲット．
HNUG*, CNUG* それぞれヘルプデスク・顧客のゴール・ナゲット．問題解決に至ったことがわかるターンを意味する．
HNaN, CNaN ナゲットではない (Not a Nugget)，すなわち，問題解決に貢献しないターン．

正解およびシステム出力に単一のラベルではなく分布を用いているのは，対話破綻検出チャレンジ (Dialogue Breakdown Detection Challenge) [4] にヒントを得ている．分布を多数決などにより単一のラベルにしてしまうと，例えば対話品質の評価が完全に割れる場合と，満場一致となる場合との違いが失われてしまう．対話システムは，同じ発話に対する受け取り方がユーザによって異なることも考慮すべきという思想から，正解分布をそのまま評価に用いるアプローチをとっている．また，DQ サブタスクによる対話全体の品質評価に加えて ND サブタスクによるターンレベルの評価を行うことは，対話中のどの部分に問題があったかを自己診断する対話システムの要素技術開発に役立つと考えている．

DQ, ND サブタスクともに，評価はシステムによる推定分布と正解分布との比較に基づき行う．しかし，両者では異なる評価指標を用いている．その理由は，ND サブタスクの分布は名義的ビン (ナゲットの種類) 上に定義されるのに対し，DQ サブタスクの分布は順序的ビン (対話品質評価値) 上に定義されるものであるからである．具体的には，ND サブタスクでは RNSS (Root Normalised Sum of Squares) および JSD (Jensen-Shannon Divergence) が，DQ サブタスクでは NMD (Normalised Match Distance) および RSNOD (Root Symmetric Normalised Order-aware Divergence) [11] という評価指標が用いられている．

図 2 は，順序尺度上の分布を扱う際の評価指標の重要性を示すために，DQ サブタスクにおける評価イメージを示したものである．この例では，システム X のほうがシステム Y より優れているという評価を下すべきであろう．前述の NMD および RSNOD は，ビン間の距離の概念を利用しているため適切な評価を下すことができる．一方，JSD などのように正解分布と推定分布をビン毎に比較しその誤差の総和をとる指標では，システム X がシステム Y と同等と見なされてしまう[*17]．

[*15] Twitter に似た中国のマイクロブログサイト．
[*16] STC-3 の第 3 のサブタスク (指定された感情にマッチする応答の生成に関するもの) については別個のオーバービュー論文があるのでこちらをご参照いただきたい [19].
[*17] 対話破綻検出チャレンジではビン毎の誤差の総和をとる評価指標 (JSD および平均二乗誤差) が採用されているが，このタスクも順序的ビン (「破綻である」「破綻かもしれない」「破綻でない」) を扱うものであるから [4]，NMD や RSNOD を用いるほうが適切であると筆者は考える．また，上記 3 つのビンを 2 つにつぶし

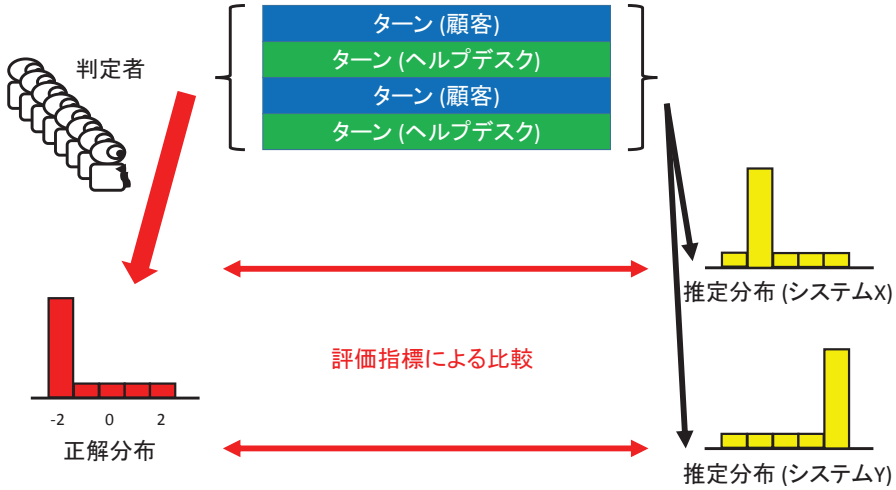

図 2 対話品質サブタスクにおける順序尺度上の分布間の比較.

3.1.2 NTCIR-14 STC-3 英語サブタスク公式結果

NTCIR-14 STC-3 における英語を対象とした各サブタスクのトップのシステムは以下の通りである．ただし，上位のシステム間に統計的有意差は見られなかった．

DQ, A-score オーガナイザが参加者に提供した Bidirectional Long Short-Term Memory (BiLSTM) に基づくシステム (`BL-lstm`) がトップであった．

DQ, S-score 早稲田大学の `BL-lstm` に改良を施したラン `SLSTC-run1` (損失関数に隣接するビンの概念を導入し，DQ と ND の同時学習を行うもの) および `SLSTC-run2` (埋め込み層の分散表現に Bidirectional Encoder Representation from Transformers (BERT) [1] を利用したもの) がトップであった [5].

DQ, E-score `SLSTC-run2` および `BL-lstm` がトップであった．

ND `BL-lstm` がトップであった．

3.2 DialEval-1 のタスク設定

DialEval-1 のタスク設定は，STC-3 における DQ および ND サブタスクと同一である．テストデータとして，新たに Weibo からマイニングした 300 件程度の対話データが提供される予定である[*18]．また，STC-3 においては中国語の学習用データ 3,700 件のうち 1,672 件のみの英訳が提供されたが，英語の学習用データの分量も増やす予定である．さらに，STC-3 の 390 件のテストデータも DialEval-1 のための学習用もしくはバリデーションデータとして活用できる．DialEval-1 参加者は，ランの提出に先立ち，オーガナイザが提供するリーダーボードサイトに上記 STC-3 のテストデータ (Dialog-1 のテストデータではない) の処理結果を提出することにより，有効な手法の選定やシステムの調整を行うことができる．対話システム，特にタスク指向の対話システムに興味のある方には是非本タスクへの参加をお願いしたい．

4. まとめ

表 1 に NTCIR-15 WWW-3 と DialEval-1 のスケジュールを示す．両タスクの参加登録締切はそれぞれ 2020 年 4 月および 6 月であり，まだたっぷり時間があるので是非ご検討いただきたい．本稿では両タスクの概要のみについて説明したが，詳細については online proceedings のタスクオーガナイザによる論文 (overview papers) および参加者による論文 (participant papers) をご参照いただきたい[*19]．なお，NTCIR-15 の全タスク参加者には 2020 年 8 月の評価結果配布を受けたタスク参加者論文の執筆と，NTCIR-15 におけるポスター発表 (もしくは口頭発表) が義務付けられている．日本在住の研究者にとっての NTCIR は，ワールドクラスの研究者達とホームグラウンドにおいて比較的気軽に議論を交わせる場なので，有効活用していただきたい．

最後に，筆者が今回紹介した両タスクにピンと来なかった読者には，NTCIR-15 の他のタスクへの参加のご検討もお勧めする．NTCIR の Twitter アカウント `@ntcir` をフォローするなどして，他のタスクの情報を収集していただきたい．これは NTCIR の元共同ジェネラルチェアとしてのお願いである．

た上で評価を行えば [14] ビン間の距離の概念すなわち順序尺度の問題が回避できるが，これは「破綻である」と「破綻かもしれない」もしくは「破綻かもしれない」と「破綻ではない」を同一視することを意味し，本質的な解決策ではないと考える．

[*18] このテストデータの件数も，STC-3 の結果をもとにトピック数設計 [12], [21] により決定した．

[*19] `http://research.nii.ac.jp/ntcir/publication1-ja.html`

表 1　NTCIR-15 WWW-3 と DialEval-1 のスケジュール (下線部は参加者のアクション).

	WWW-3		DialEval-1
		2019 年 7 月	テストデータのクロール
		2019 年 8-10 月	訓練用データの英訳追加
2019 年 10 月	タスク登録・CENTRE 実験開始	2019 年 10 月	タスク登録開始
		2019 年 10-12 月	テストデータのアノテーション
2020 年 2 月	WWW-3 トピック集合公開		
2020 年 4 月	タスク登録締切		
2020 年 5 月	結果提出締切		
2020 年 6-7 月	適合性判定	2020 年 6 月	タスク登録締切・テストデータ公開
		2020 年 7 月	結果提出締切
2020 年 8 月	評価結果配布		
2020 年 12 月	NTCIR-15 カンファレンス (東京)，Springer の NTCIR 本を参加者に配布		
2021 年 3 月	Post-conference proceedings 発行		

謝辞　NTCIR の各チェア，プログラム委員の皆様，WWW, CENTRE, STC, DialEval-1 共同オーガナイザの皆様，そして WWW, CENTRE, STC に参加してくださった皆様に感謝します.

参考文献

[1] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (online), available from ⟨http://arxiv.org/abs/1810.04805⟩ (2018).
[2] Fang, H. and Zhai, C.: Semantic Term Matching in Axiomatic Approaches to Information Rerieval, *Proceedings of ACM SIGIR 2006*, pp. 115–122 (2006).
[3] Gallagher, L., Mackenzie, J., Benham, R., Chen, R.-C., Scholer, F. and Culpepper, J. S.: RMIT at the NTCIR-13 We Want Web Task, *Proceedings of NTCIR-13*, pp. 402–406 (2017).
[4] Higashinaka, R., D'Haro, L. F., Shawar, B. A., Banchs, R. E., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T. and ao Sedoc, J.: Overview of the Dialogue Breakdown Detection Challenge 4, *Proceedings of Chatbots and Conversational Agents and Dialogue Breakdown Detection Challenge (WOCHAT+DBDC), IWSDS 2019* (2019).
[5] Kato, S., Suzuki, R., Zeng, Z. and Sakai, T.: SLSTC at the NTCIR-14 STC-3 Dialogue Quality and Nugget Detection Subtasks, *Proceedings of NTCIR-14*, pp. 355–361 (2019).
[6] Lin, J.: The Neural Hype and Comparisons against Weak Baselines, *SIGIR Forum*, Vol. 52, pp. 40–51 (2019).
[7] Luo, C., Sakai, T., Liu, Y., Dou, Z., Xiong, C. and Xu, J.: Overview of the NTCIR-13 We Want Web Task, *Proceedings of NTCIR-13*, pp. 394–401 (2017).
[8] Mao, J., Sakai, T., Luo, C., Xiao, P., Liu, Y. and Dou, Z.: Overview of the NTCIR-14 We Want Web Task, *Proceedings of NTCIR-14*, pp. 455–467 (2019).
[9] Metzler, D. and Croft, W. B.: A Markov Random Field Model for Term Dependencies, *Proceedings of ACM SIGIR 2005*, pp. 472–479 (2005).
[10] Qin, T. and Liu, T.-Y.: Introducing LETOR 4.0 Datasets, (online), available from ⟨http://arxiv.org/abs/1306.2597⟩ (2013).
[11] Sakai, T.: Comparing Two Binned Probability Distributions for Information Access Evaluation, *Proceedings of ACM SIGIR 2018*, pp. 1073–1076 (2018).
[12] Sakai, T.: *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*, Springer (2018).
[13] Sakai, T., Ferro, N., Soboroff, I., Zeng, Z., Xiao, P. and Maistro, M.: Overview of the NTCIR-14 CENTRE Task, *Proceedings of NTCIR-14*, pp. 494–509 (2019).
[14] Tsunomori, Y., Higashinaka, R., Takahashi, T. and Inaba, M.: Evaluating Dialogue Breakdown Detection in Chat-Oriented Dialogue Systems, *Proceedings of SEMDIAL 2018* (2018).
[15] Yang, P. and Fang, H.: Evaluating the Effectiveness of Axiomatic Approaches in Web Track, *Proceedings of TREC 2013* (2014).
[16] Yates, A.: MPII at the NTCIR-14 CENTRE Task, *Proceedings of NTCIR-14*, pp. 510–513 (2019).
[17] Zeng, Z., Kato, S. and Sakai, T.: Overview of the NTCIR-14 Short Text Conversation Task: Dialogue Quality and Nugget Detection Subtasks, *Proceedings of NTCIR-14*, pp. 289–315 (2019).
[18] Zeng, Z., Luo, C., Shang, L., Li, H. and Sakai, T.: Towards Automatic Evaluation of Customer-Helpdesk Dialogues, *Journal of Information Processing*, Vol. 26, pp. 768–778 (2018).
[19] Zhang, Y. and Huang, M.: Overview of the NTCIR-14 Short Text Generation Subtask: Emotion Generation Challenge, *Proceedings of NTCIR-14*, pp. 316–327 (2019).
[20] Zheng, Y., Chu, Z., Li, X., Mao, J., Liu, Y., Zhang, M. and Ma, S.: THUIR at the NTCIR-14 WWW-2 Task, *Proceedings of NTCIR-14*, pp. 472–480 (2019).
[21] 酒井哲也: 情報アクセス評価方法論: 検索エンジンの進歩のために，コロナ社 (2015).