Vol.2021-DBS-173 No.10
Vol.2021-IFAT-144 No.10
2021/9/17

NTCIR-16 ウェブ検索・再現可能性タスク (WWW-4) および対話評価タスク (DialEval-2) への誘い

酒井 哲也[1,a)]

概要：本稿の目的は，NTCIR-16 で開催される We Want Web with CENTRE (WWW-3) タスクおよび Dialogue Evaluation (DialEval-2) 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-4 は英語ウェブ検索タスクであり，このために新しいウェブコーパスを構築中である．また，参加チームが NTCIR-15 WWW-3 においてトップであったシステムを再現できるかという課題も提供している．一方，DialEval-2 は NTCIR-14 Short Text Conversation および NTCIR-15 DialEval-1 と同様に中国語もしくは英語の顧客・ヘルプデスク間の対話の品質を扱うタスクであり，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，タスク参加者に何が求められるかを説明する．

Invitation to the NTCIR-16 We Want Web with CENTRE (WWW-4) and Dialogue Evaluation (DialEval-2) Tasks

Tetsuya Sakai[1,a)]

Abstract: The objective of this paper is to try to convince the reader to participate in the NTCIR-16 We Want Web with CENTRE Task (WWW-4) and/or the Dialogue Evaluation Task (DialEval-2). WWW-3 is an ad hoc English web search task, and we are currently constructing a new target web corpus for it. Participants may also try to reproduce the top run from the NTCIR-15 WWW-3 Task. On the other hand, DialEval-2 is a continuation of the NTCIR-14 Short Text Conversation Task and the NTCIR-15 DialEval-1 Task, and addresses the quality of customer-helpdesk dialogues written in Chinese or English. 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 previous rounds and a description of what is expected of a 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 [14], [23], 2022 年 6 月には第 16 回 (NTCIR-16) を迎える．本稿では，NTCIR-16 で開催される第 4 回 We Want Web with CENTRE タスク (WWW-4)[*2] および第 2 回 Dialogue Evaluation タスク (DialEval-1)[*3] 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-16 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．なお，筆者は NTCIR-15 に向けた同様の主旨の研究報告をちょうど 2 年前に行っている [24].

[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/www4/
[*3] http://sakailab.com/dialeval2

表 1　WWW タスクの提出ラン数 (および参加チーム数).

	Chinese	English	Chinese ∪ English
WWW-1	19 (4)	13 (3)	32 (5)
WWW-2	11 (3)	20 (5)	31 (5)
WWW-3	11 (3)	37 (9)	48 (9)

2. We Want Web with CENTRE (WWW-4) タスクへの誘い

2.1 WWW タスクの歩み

TREC 2014[*4]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[*5]．ウェブ検索は「解決済の研究課題」ではないためであろう．現存の検索エンジンはキーワードマッチングを基本としており，ユーザの情報要求に必ずしも的確に答えてくれるものではない．また，ユーザからは見えないが，適合する情報を検索エンジンが検索しそこねているかも知れない．さらに，近年，深層学習が検索結果の再ランキングに利用され始めているが，このような新しいアプローチはまた新たな研究課題を生みだしている．繰り返すが，ウェブ検索は「解決済の研究課題」ではない．

上記の観点から，筆者らは，2017 年に NTCIR-13 において中国語サブタスクと英語サブタスクから構成される We Want Web (WWW-1) タスクを開催した [5]．これは，検索対象ウェブコーパスと評価用トピック (検索課題) 集合が与えられたときに，各トピックに対するランクつき検索結果を出力するという典型的なアドホック検索タスク [17], [23] である．2019 年の NTCIR-14 WWW-2 タスク [6] もタスク仕様は同じであった．一方，2020 年の NTCIR-15 WWW-3 タスク [13] は，他の研究機関が報告した実験結果の再現可能性 (reproducibility) を検証することを目的とした NTCIR-14 CENTRE タスク [12] と合流したタスクとなった．CENTRE は **CLE**F **N**TCIR **TRE**C Reproducibility の略で，もともとはアジア母体の NTCIR，欧州母体の CLEF[3]，および米国母体の TREC [17] をまたがるプロジェクトであった．これにより，現在の WWW タスクの主要なミッションは以下の 2 つである．

(1) ウェブ検索技術の進歩を定量化する．
(2) SOTA (state-of-the-art) なウェブ検索技術の再現可能性を検証する．

上記の両輪がうまく回って初めて，他者の研究成果の上に自分の研究成果を積み上げていくことによるコミュニティとしての技術進歩が可能となると考える．

表 1 に，WWW タスクに提出されたラン (検索結果ファイル) 数と参加チーム数をまとめた．中国語サブタスクは参加チーム数が低迷しているため，WWW-4 ではこれを廃止し，英語サブタスクのみ継続することとした．

2.2 WWW-4 タスクの内容

WWW-4 (英語のみ) のタスク仕様は従来と同様，与えられた評価用トピック集合の各トピックに対しウェブページの文書 ID のランクつきリストを出力したランファイルを提出するというものである．詳細については WWW-4 のウェブページをご覧いただきたい.[*6] 従来の英語サブタスクとの差分は，検索対象コーパスとして `clueweb12-B13`[*7]ではなく，Common Crawl データ[*8] をもとに構築した独自のコーパスを用いる点である．Clueweb が 2012 年にクロールされたデータであるのに対し，我々の新コーパスは 2021 年にクロールされたものであるので，今後，検索タスクに限らず様々な研究用途で活用されることを期待している.[*9]

評価用として 50 件のトピックが 10 月に公開される.[*10]各参加チームが提出できるランには以下の 3 種類がある．

REV (revised) run　このランは WWW-3 英語サブタスクにおいて最も成績のよいランを提出した筑波大学の KASYS チーム [16] のみが提出可能である．具体的には，KASYS チームが WWW-3 タスクに提出したラン `KASYS-E-CO-NEW1` の作成に用いたシステムを一切変更せずに，WWW-4 のトピック集合に適用してもらう．このランは Yilmaz らによる BERT[*11]ベースの手法 [18] を利用したもので，WWW-3 の公式結果における nDCG，Q-measure，および iRBU [15] という 3 つの評価指標においてトップの成績を示した [13]．従って，この手法を WWW-3 タスク終了時点での SOTA と見なす．

NEW runs　SOTA の更新を目指す通常の参加チームのランである．オーガナイザおよび参加者は，NEW runs を上記 REV run と同じテストコレクション上で統計的見地から比較することにより，「技術進歩」が実質的なものであるかどうか議論できる．なお，参加チームは，オーガナイザが提供する BM25[7] に基づくベースラインランを再ランキングしただけのランを提出することも可能である．すなわち，必ずしも検索対象コーパス全体を索引づけしなくても参加できる．

REP (reproduced) runs　こちらは，検索技術の再現可能性に興味があるチームが提出するランである．具体的には，上記 KASYS チームの論文および Yilmaz らの論文を読んで，`KASYS-E-CO-NEW1` の手法を独自に

[*4] `https://trec.nist.gov/pubs/trec23/trec2014.html`
[*5] `https://twitter.com/djoerd/status/536128465276530688`
[*6] `http://sakailab.com/www4/`
[*7] `http://lemurproject.org/clueweb12/`
[*8] `http://commoncrawl.org/`
[*9] 新コーパス構築のためにご尽力されている清華大学の Zhumin Chu 君と Yiqun Liu 先生に感謝する．
[*10] この件数は，過去の評価結果から各評価指標の母分散を推定し，これをトピック数設計ツール [10] に投入することにより統計的見地から決定したものである．
[*11] Bidirectional Encoder Represenations from Transformers [2]

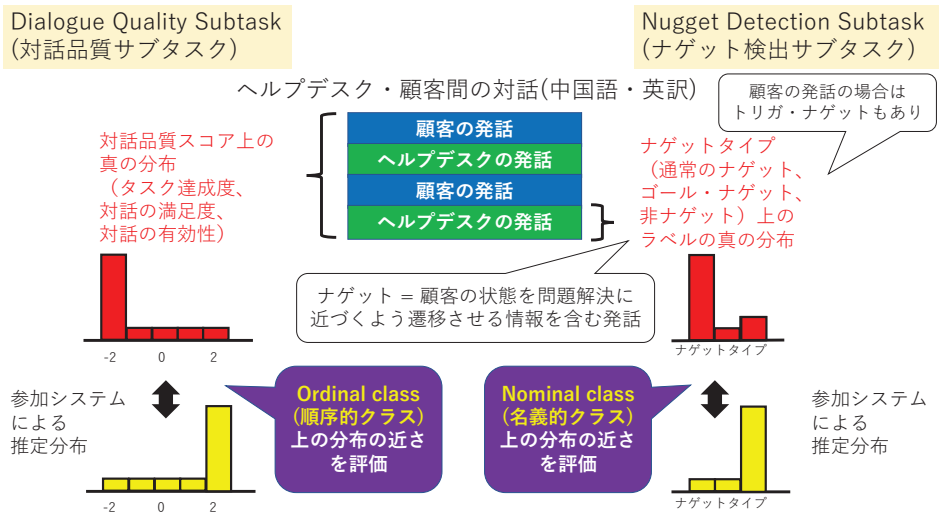

図 1 DialEval タスク概観.

表 2 DialEval タスクの提出ラン数 (および参加チーム数).

	Dialogue Quality			Nugget Detection		
	Chinese	English	Chinese ∪ English	Chinese	English	Chinese ∪ English
STC-3	7 (3)	6 (3)	13 (4)	6 (2)	5 (2)	11 (3)
DialEval-1	10 (5)	6 (4)	16 (6)	12 (6)	6 (3)	18 (6)

再現し，そのシステムを WWW-4 のトピック集合に適用したものである．再現可能性の評価 [1] は，REP runs を REV run と比較することにより行う．

情報検索に興味のある研究者の方々 (特に学生の方々) には，このタスクを利用してグローバルな舞台で検索技術の進歩に貢献していただきたい．

3. Dialogue Evaluation (DialEval-2) タスクへの誘い

3.1 Dialogue Evaluation タスクの歩み

図 1 に Dialogue Evaluation Task の概観を示す．その目的は，顧客・ヘルプデスク間の対話における問題解決過程と対話品質についての知見を深め，将来的にはそれを多様な顧客ニーズに対応する自動ヘルプデスクシステムの構築に役立てることである．Weibo[*12]から収集した中国語のヘルプデスク対話データ，およびこれらを人手により英訳したデータを扱うものであり，タスク仕様は NTCIR-14 Short Text Conversation Task [19] 以来変わっていない．具体的には，以下の 2 つのサブタスクにより構成される．

対話品質 (Dialogue Quality: DQ) オーガナイザは，評価用の各ヘルプデスク対話について，20 名程度の判定者から 5 段階の対話品質評価スコア (後述) を収集し集計しておく．参加チームは，各対話について，このスコアの分布を予測する．

ナゲット検出 (Nugget Detection: ND) DialEval では，問題を抱えている顧客の状態が問題解決に近づくよう遷移させる情報を含む発話をナゲットと呼んでいる．オーガナイザは，上記対話中のヘルプデスクの各発話および顧客の各発話について，20 名程度の判定者からナゲットタイプ (後述) のラベルを収集し，集計しておく．参加チームは，各ヘルプデスク対話中の各発話について，このラベルの分布を予測する．

DQ サブタスクに対す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 対話の満足度 (customer **S**atisfaction)
E-score 対話の有効性 (dialogue **E**ffectiveness)

各判定者はそれぞれ独立に，−2 点から 2 点までの 5 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発話は，各判定者によりそれぞれ独立に以下のいずれか 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発話を意味する．
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発話を意味する．
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発話．

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) [22] にヒントを得たもの

[*12] https://weibo.com/

表 3 NTCIR-16 WWW-4 と DialEval-2 のスケジュール (下線部は参加者のアクション).

	WWW-4		DialEval-2
2021 年 10 月 1 日	評価用トピック集合公開，タスク登録締切		
2021 年 11 月 15 日	結果提出締切		
		2021 年 12 月 1 日	評価用対話集合公開，タスク登録締切
2021 年 12 月〜2022 年 1 月	適合性判定	2022 年 1 月 15 日	結果提出締切
2022 年 2 月 1 日	評価結果および overview 論文の初稿配布		
2022 年 3 月 1 日	参加者論文初稿締切		
2022 年 5 月 1 日	全ての論文の camera-ready 原稿締切		
2022 年 6 月 14〜17 日	NTCIR-16 カンファレンス (東京・ハイブリッド)		

である．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，同じシステム発話に対する受け取り方がユーザによって異なる可能性も考慮して対話戦略を立てるべき」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 サブタスクによる発話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 サブタスクの分布が nominal class (名義的クラス) であるナゲットタイプ上に定義されるのに対し，DQ サブタスクの分布は ordinal class (順序的クラス) である対話品質評価スコア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．本質的には Root Mean Squared Error と同じ) および JSD (Jensen-Shannon Divergence) が，DQ サブタスクでは NMD (Normalised Match Distance．本質的には Earth Mover's Distance と同じ) および RSNOD (Root Symmetric Normalised Order-aware Divergence) [9] という評価指標が用いられている．

DQ サブタスクのように，順序的クラス上の分布を予測するタスクを ordinal quantification という．筆者は，SemEval の ordinal quantification タスクのデータ [8] および NTCIR-15 DialEval-1 [20] のタスクデータを用いた実験において，RNOD (Root Normalised Order-aware Divergence．前述の RSNOD から対称性の性質を除いたもの) のほうが NMD よりも統計的に好ましい性質を有していることを示した [11].[*13] 角森ら [22] も，対話破綻検出チャレンジのデータを用いた実験にもとづき，同様の統計的観点から RSNOD のほうが NMD よりも総合的に好ましい性質を示したと報告している．

表 2 にこれまでに DQ および ND サブタスクに提出されたラン数と参加チーム数を示す．

3.2 DialEval-2 タスクの内容

DialEval-2 参加チームは，訓練用データとして，我々がこれまでに構築したヘルプデスク対話データをまとめた DCH-2 データセット [21] を利用することができる．DCH-2 には DQ，ND サブタスクのラベルを含む中英それぞれ 4,390 件のヘルプデスク対話が収録されている．評価用の対話としては，新たに 66 件の対話が配布される予定である.[*14]

従来と同様に，オーガナイザから LSTM [*15]ベースのベースラインシステムなどを提供する予定なので，このシステムを調整する形で参加することも可能である．なお，NTCIR-15 DialEval-1 では，BERT ベースのアプローチにより LSTM ベースラインを統計的に有意に上回る成績を残した参加チームが複数登場した [20]．さらに，英語データのみ，もしくは中国語データのみを対象とした参加や，DQ，ND いずれかのサブタスクのみの参加も可能なので，ご検討いただきたい．

4. まとめ

表 3 に NTCIR-16 WWW-4 と DialEval-2 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 2021 年 10 月 1 日および 12 月 1 日である．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 NTCIR online proceedings[*16]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．なお，NTCIR-16 の全タスク参加者には，タスク参加者論文の執筆と NTCIR-16 におけるポスター発表 (もしくは口頭発表) が義務づけられている．オ

[*13] この論文の日本語の解説記事を発行予定である [25]．また，この論文に付随するコンテンツとして，順序的クラスを扱う ordinal classification および ordinal quantification タスクのための評価指標に関する英語のチュートリアルをこちらに用意した．https://waseda.box.com/acl2021sakai-videos-and-slides

[*14] このサンプルサイズも，DialEval-1 の評価結果 [20] から得た各評価指標の母分散の推定値と，トピック数設計ツール [10] により決定したものである．

[*15] Long Short-Term Memory [4]

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

ンライン参加も可能である．日本在住の研究者にとってのNTCIRは，ホームグラウンドにいながらワールドクラスの研究者達と気軽に議論を交わせる場なので，積極的に活用していただきたい．

謝辞 NTCIRの各チェア，プログラム委員の皆様，WWWタスクとDialEvalタスクの共同オーガナイザの皆様 (とくにWWW-4とDialEval-2のオーガナイザであるZhumin Chu氏，Nicola Ferro氏，Inho Kang氏，Yiqun Liu氏，Maria Maistro氏，Ian Soboroff氏，Sijie Tao氏)，そしてこれまで各タスクに参加してくださった皆様 (とくにWWW-4において再現可能性のターゲットとさせていただいた筑波大学KASYSチームの皆様) に感謝します．

参考文献

[1] Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P. and Soboroff, I.: How to Measure the Reproducibility of System-oriented IR Experiments, *Proceedings of ACM SIGIR 2020*, pp. 349–358 (2020).

[2] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *CoRR*, Vol. abs/1810.04805 (online), available from ⟨http://arxiv.org/abs/1810.04805⟩ (2018).

[3] Ferro, N. and (eds.), C. P.: *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, Springer (2019).

[4] Goodfellow, I., Bengio, Y. and Courville, A.: *Deep Learning*, The MIT Press (2016).

[5] 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).

[6] 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).

[7] Robertson, S. and Zaragoza, H.: The Probabilistic Relvance Framework: BM25 and Beyond, *Foundations and Trends in Information Retrieval*, Vol. 3, No. 4 (2009).

[8] Rosenthal, S., Farra, N. and Nakov, P.: SemEval-2017 Task 4: Sentiment Analysis in Twitter, *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, Association for Computational Linguistics, (online), DOI: 10.18653/v1/S17-2088 (2017).

[9] Sakai, T.: Comparing Two Binned Probability Distributions for Information Access Evaluation, *Proceedings of ACM SIGIR 2018*, pp. 1073–1076 (2018).

[10] Sakai, T.: *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*, Springer (2018).

[11] Sakai, T.: Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification, *Proceedings of ACL-IJCNLP 2021* (2021).

[12] 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).

[13] Sakai, T., nad Zhaohao Zeng, S. T., Zheng, Y., Mao, J., Chu, Z., Liu, Y., Maistro, M., Dou, Z., Ferro, N. and Soboroff, I.: Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task, *Proceedings of NTCIR-15*, pp. 219–234 (2020).

[14] Sakai, T., Oard, D. W. and (eds.), N. K.: Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact, Springer, (オンライン), 入手先 ⟨https://link.springer.com/content/pdf/10.1007%2F978-981-15-5554-1.pdf⟩ (2020).

[15] Sakai, T. and Zeng, Z.: Retrieval Evaluation Measures that Agree with Users' SERP Preferences: Traditional, Preference-based, and Diversity Measures, *ACM TOIS*, Vol. 39, No. 2 (2021).

[16] Shinden, K., Maruta, A. and Kato, M. P.: KASYS at the NTCIR-15 WWW-3 Task, *Proceedings of NTCIR-15*, pp. 235–238 (2020).

[17] Voorhees, E. M. and Harman, D. K.: *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press (2005).

[18] Yilmaz, Z. A., Yang, W., Zhang, H. and Lin, J.: Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval, *Proceedings of EMNLP 2019*, pp. 3490–3496 (2019).

[19] 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).

[20] Zeng, Z., Kato, S., Sakai, T. and Kang, I.: Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task, *Proceedings of NTCIR-15*, pp. 13–34 (2020).

[21] Zeng, Z. and Sakai, T.: DCH-2: A Parallel Customer-Helpdesk Dialogue Corpus with Distributions of Annotators' Labels, *CoRR*, Vol. abs/2104.08755 (online), available from ⟨https://arxiv.org/abs/2104.08755⟩ (2021).

[22] 角森唯子，東中竜一郎，高橋哲朗，稲葉通将：対話破綻検出チャレンジ3における対話破綻検出の評価尺度の選定，人工知能学会論文誌，Vol. 35, No. 1 (2020).

[23] 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).

[24] 酒井哲也：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，情報処理学会研究報告 2019-IFAT-136(20) (2019).

[25] 酒井哲也：Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification, 自然言語処理, Vol. 12 月号 (2021).