



SIGIR2017参加 (開催) 報告

～会議概要とニューラルネットな情報検索～

櫛 惇志 (東京工業大学)

自己紹介

- 氏名
 - 櫻 惇志 (けやき あつし)
- 専門分野
 - 情報検索, データ工学
 - 高精度検索, 高速検索
 - (自然言語処理)
 - 語義曖昧性解消, クエリ分析



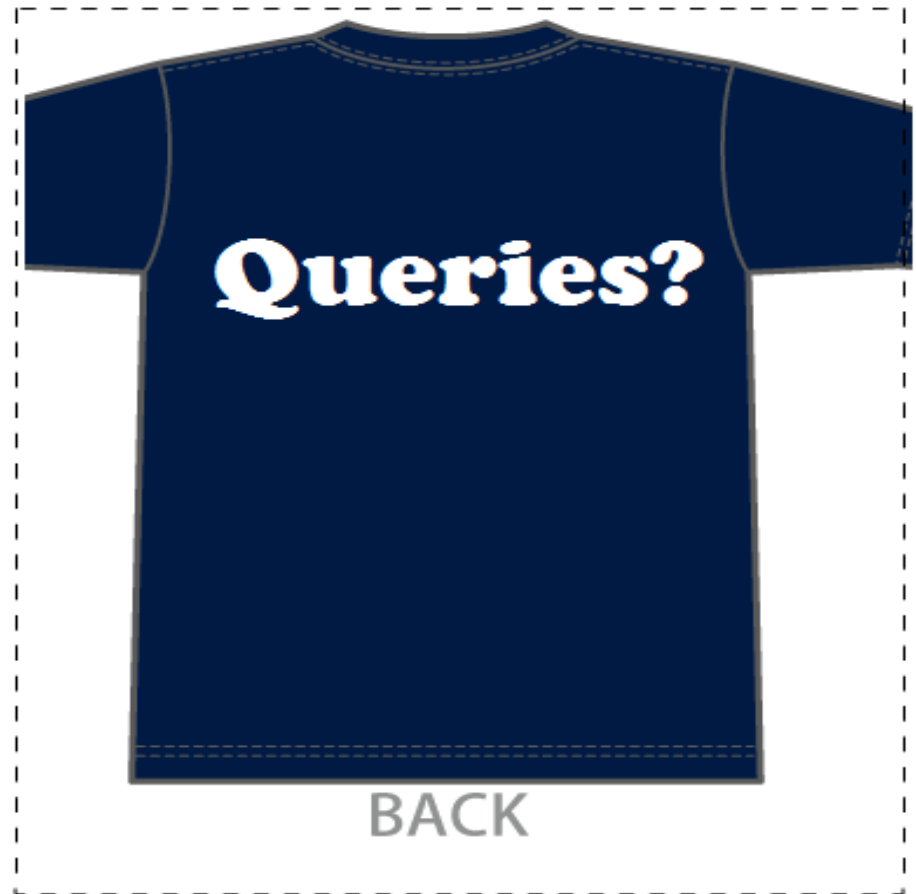
略歴

- 2011.3-2014.3 奈良先端科学技術大学院大学
博士後期課程卒業 博士(工学) 取得
- 2013.8-2014.11 **マイクロソフト・リサーチアジア** ←
研究インターン(メンター: 荒瀬先生)
- 2012.4-2014.3 日本学術振興会 特別研究員 (DC2)
- 2014.4- 東京工業大学 情報理工学研究科 (現情報理工学院) 助教着任 現在に至る
- 2016.9-2017.2 シンガポール国立大学 客員研究員

NLP との
出会い

役職	氏名
ジェネラルチェア	Noriko Kando, National Institute of Informatics
	Tetsuya Sakai, Waseda University
	Hideo Joho, University of Tsukuba
デモチェア	Takehiro Yamamoto, Kyoto University
SIRIP チェア	Sumio Fujita, Yahoo! JAPAN
広報・ソーシャル	Yukino Baba, Kyoto University
	Harumi Murakami, Osaka City University
博士コンソーシアム	Makoto P. Kato, Kyoto University
予稿集チェア	Takuya Kida, Hokkaido University
	Masaharu Yoshioka, Hokkaido University
スポンサーチェア	Tsuneaki Kato, The University of Tokyo
会計	Akiko Aizawa, National Institute of Informatics
	Hiroaki Ohshima, Kyoto University
ローカルアレンジメント	Yuka Egusa, National Institute for Educational Policy Research
	Emi Ishita, Kyushu University
	Masao Takaku, University of Tsukuba
● 学生ボランティアチェア	Atsushi Keyaki, Tokyo Institute of Technology

スタッフ T-シャツ



目次

- 国際会議紹介
- SIGIR 2017 概要
- トピック分析
- Award papers 概要
- ソーシャルプログラム写真
- Keynote 概要
 1. Stephen E. Robertson 氏
 2. Yoelle Maarek 氏
- Tutorial/Workshop
- 参加者アンケート要約
- (時間が余れば) 論文概要紹介 × 3

※ IBM Research - Tokyo camera club
(榎様, 井上様, 中澤様) ご提供と
インターネット上から取得した
写真を含む

IR 系国際会議

- ACM SIGIR 主催の国際会議
 - SIGIR (シグアイアール)
 - Special Interest Group on Information Retrieval
 - ICTIR (イクティア)
 - International Conference on the Theory of Information Retrieval
 - CHIIR (チェアー)
 - Conference Human Information Interaction and Retrieval
 - 後者二つはまだ2回しか開催されていない
- 主要国際会議・ワークショップ
 - 1st tire: The Web Conference (旧 WWW), WSDM (ウィズダム)
 - 2nd tire: CIKM (シクム, シーアイケーエム), ECIR, JCDL
 - 世界3大IR評価のWS
 - TREC (トレック), NTCIR (エンティサイル), CLEF (クレー)
 - CoNLL や SemEval と同様に shared task 開催
 - NTCIR は NII (SIGIR 2017 GC 神門先生) 主催
 - AIRS: アジア開催. 難易度は手頃
 - 情報推薦: RecSys (レクシス)

余談: DB 系国際会議

- 御三家
 - SIGMOD (シグモド)
 - VLDB (ブイエルディービー)
 - VLDB 2020 は日本開催
 - DB のラスボス退官の年
 - ICDE (アイシーディーイー)
- 主要国際会議・ワークショップ
 - EDBT
 - DASFAA (ダスファー)
 - IEEE BigData
 - IEEE BigData Congress
 - DEXA (デクサ)
 - APWeb
 - 国内会議 DEIM (デイム) とよく似たトピック

なぜか DB コミュニティでは
IR トピックでも DB 系会議に投稿☹

SIGIR 2017 概要

- 日程

7 日	チュートリアル
8-10 日	本会議
11 日	ワークショップ



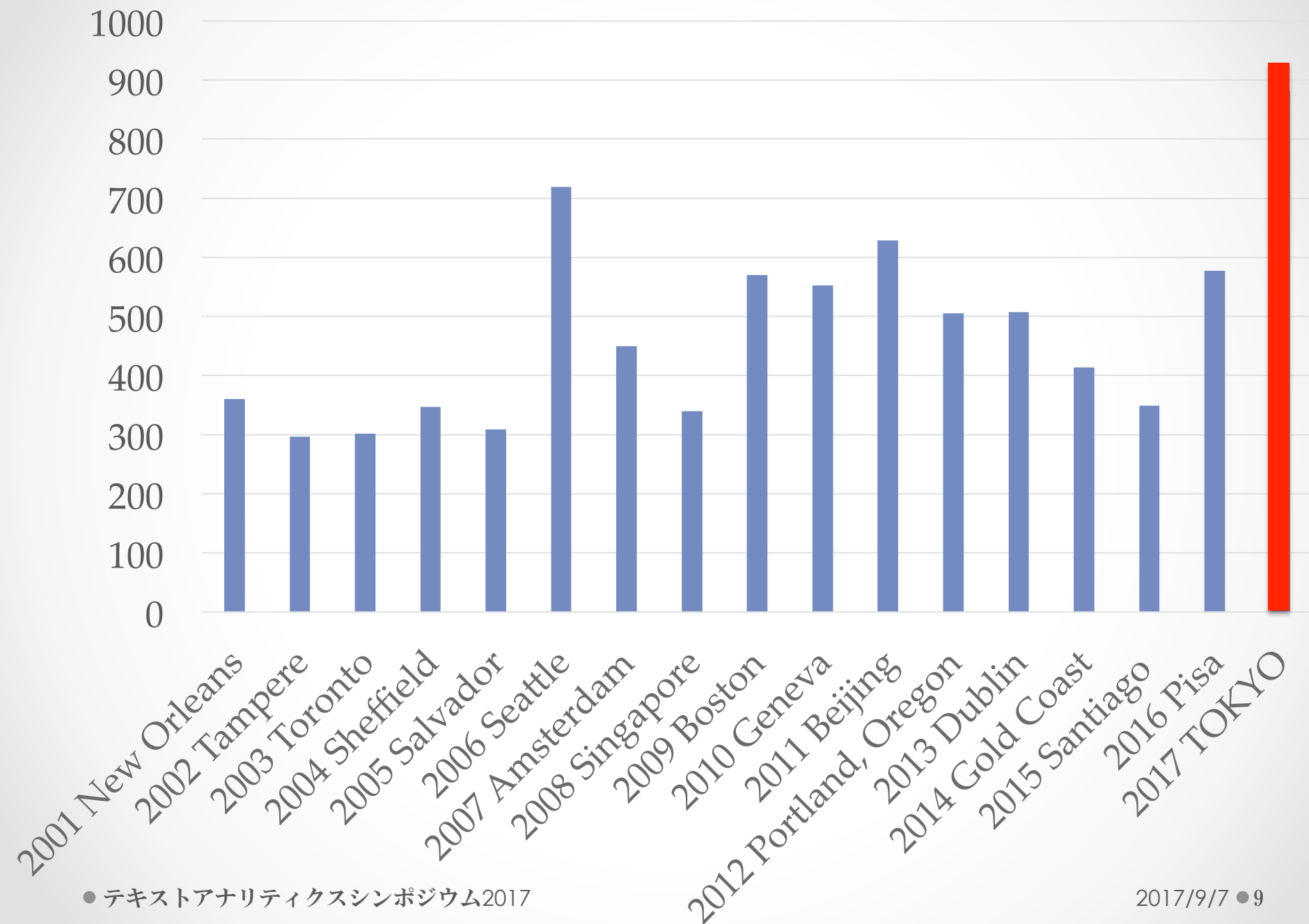
- 会議会場

- 京王プラザホテル

- 参加者

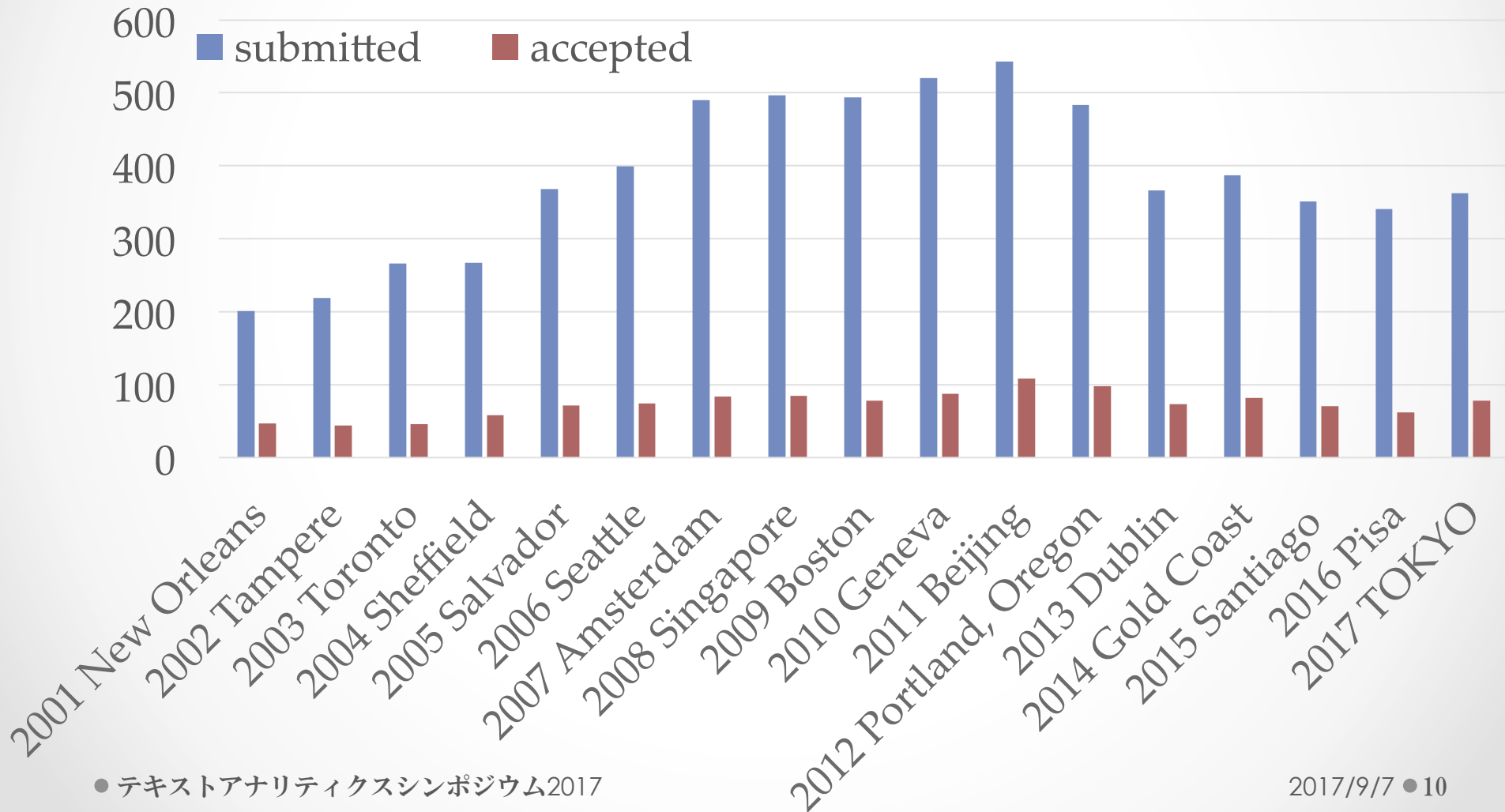
- 911 人 (過去最大)
- 国内は企業からの参加者多数
 - ・ IR も AI, ML と勘違い (?)
- NLP の方も多数観測☺
- DB 人 (特にアカポス) はあんまり☹

	人数
合計	911
日本	264
USA	177
中国	140
学生	254



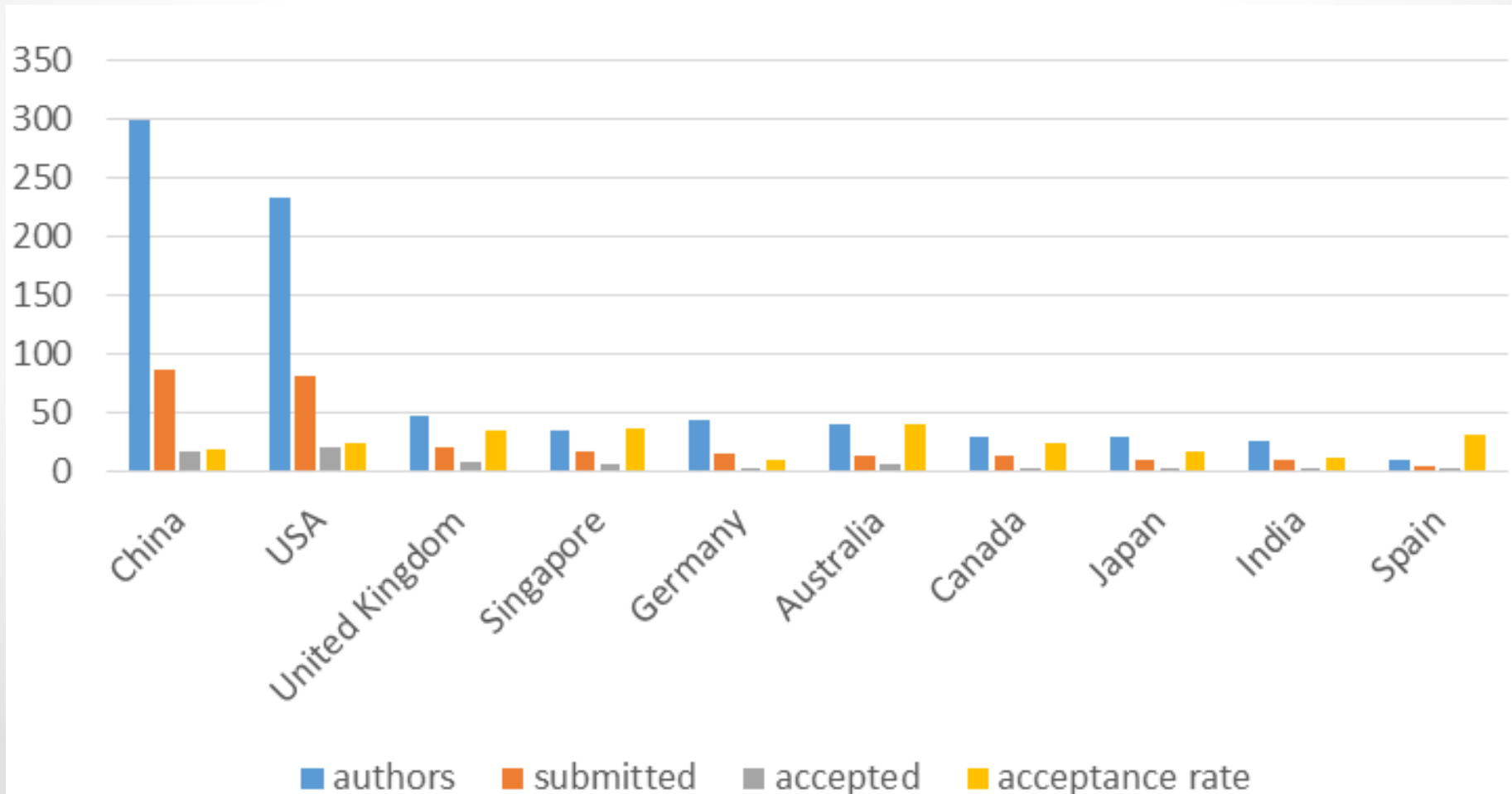
Full paper 統計量

投稿数	362
採択数	78
採択率	22%

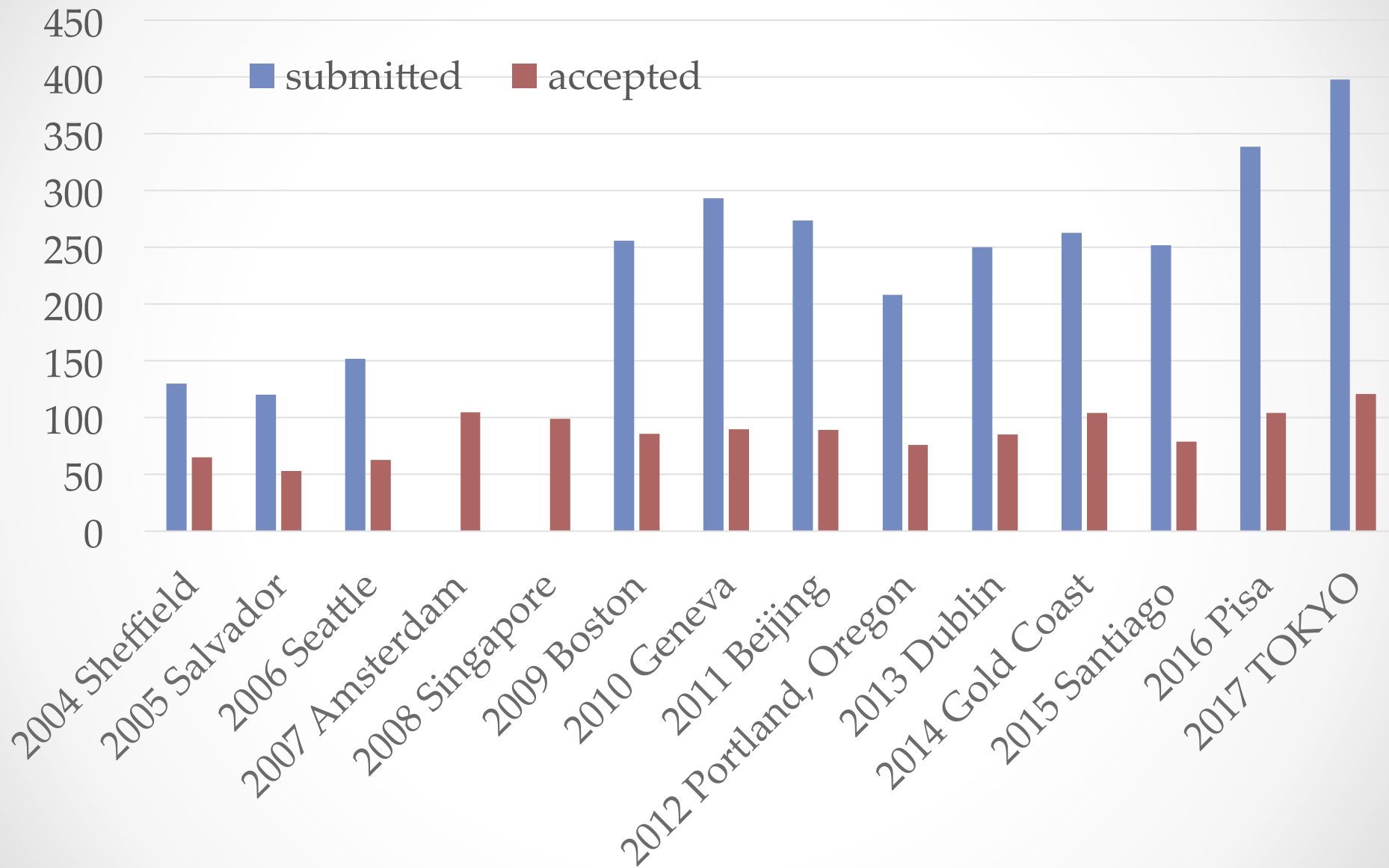


Full paper 統計量

- 投稿上位国
 - 中国, USA
- 日本は 8 位 (10 本くらい? 採択率 16%)



Short paper 統計量



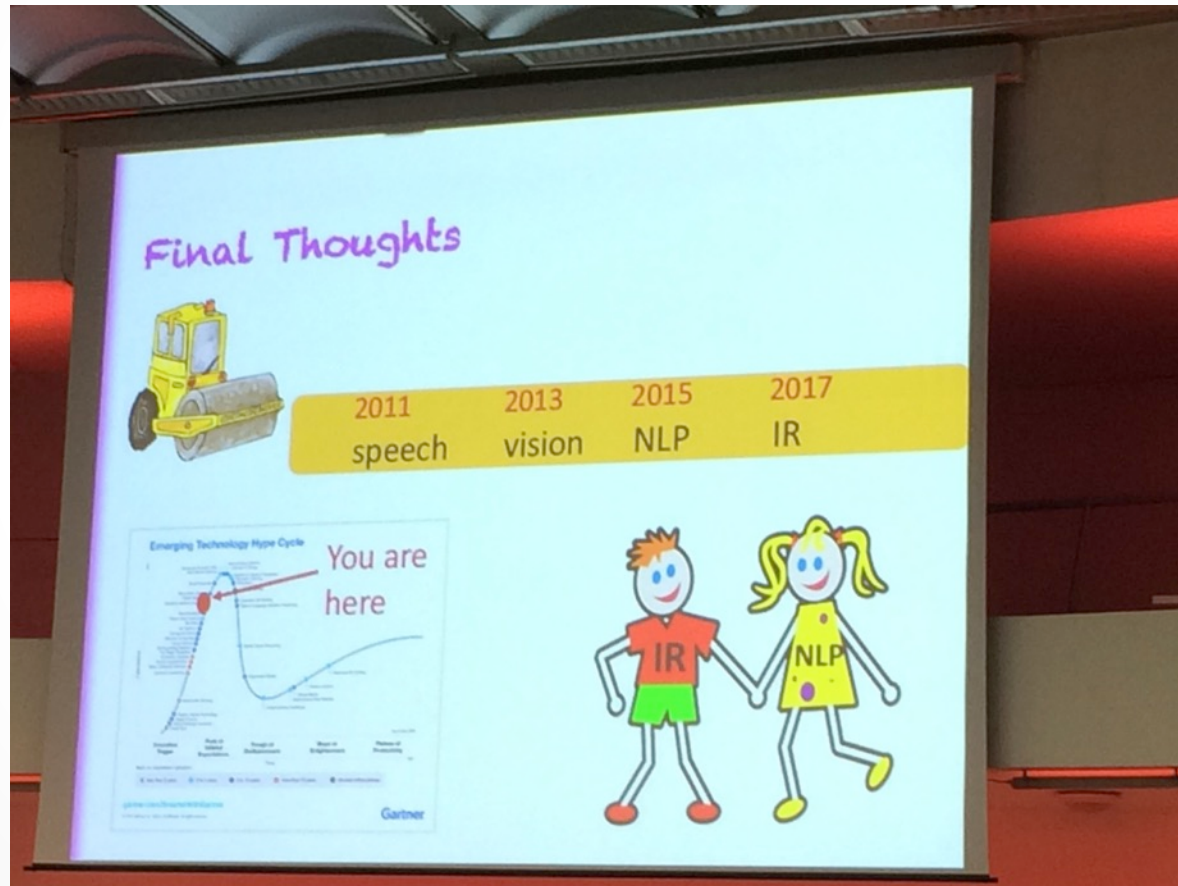
投稿トピック

どんなトピックが
流行りだと思いますか？

ACL では IR トピックが多かったと聞きました

Manning 先生のご講演@SIGIR 2016

- “2017 年に IR で深層学習のブレイクスルーがある”



投稿トピック

採択論文の 29% が
NN 技術利用

推薦が人気

NLP・ML
関連技術多数

実際 NN
大人気

推薦
システム

検索
モデル

パーソナ
ライズ

推薦アル
ゴリズム

ランキング
学習

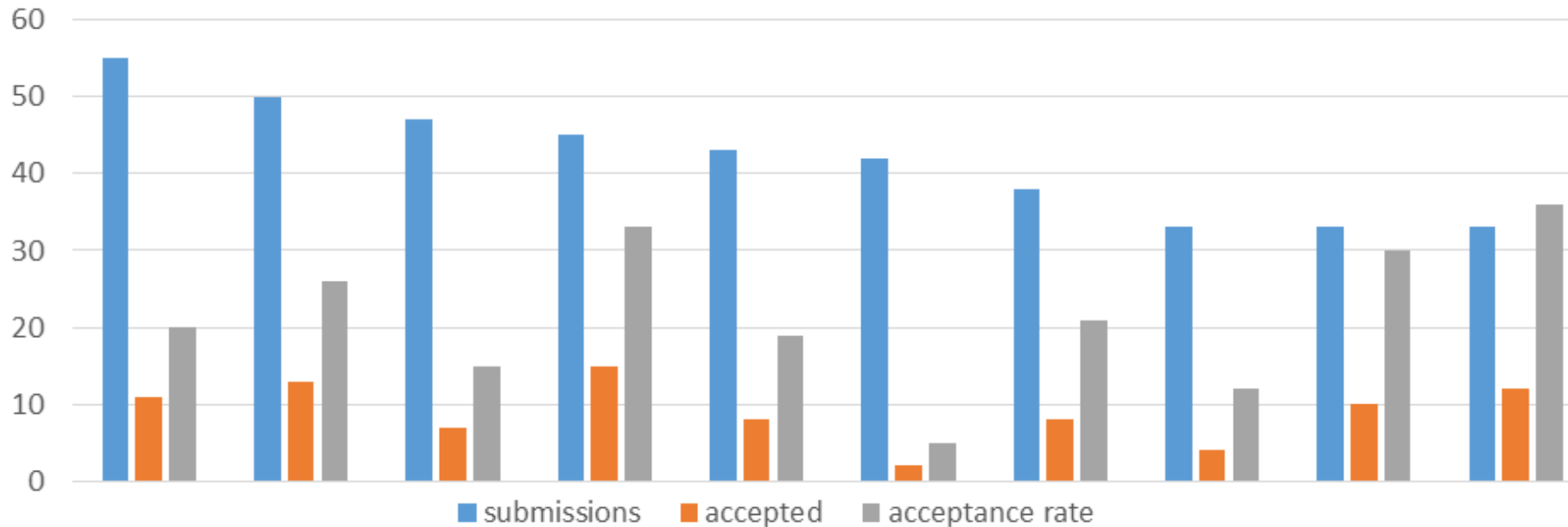
テキスト
の表現

ランキング
アルゴリズム

情報
抽出

新しい
検索タスク

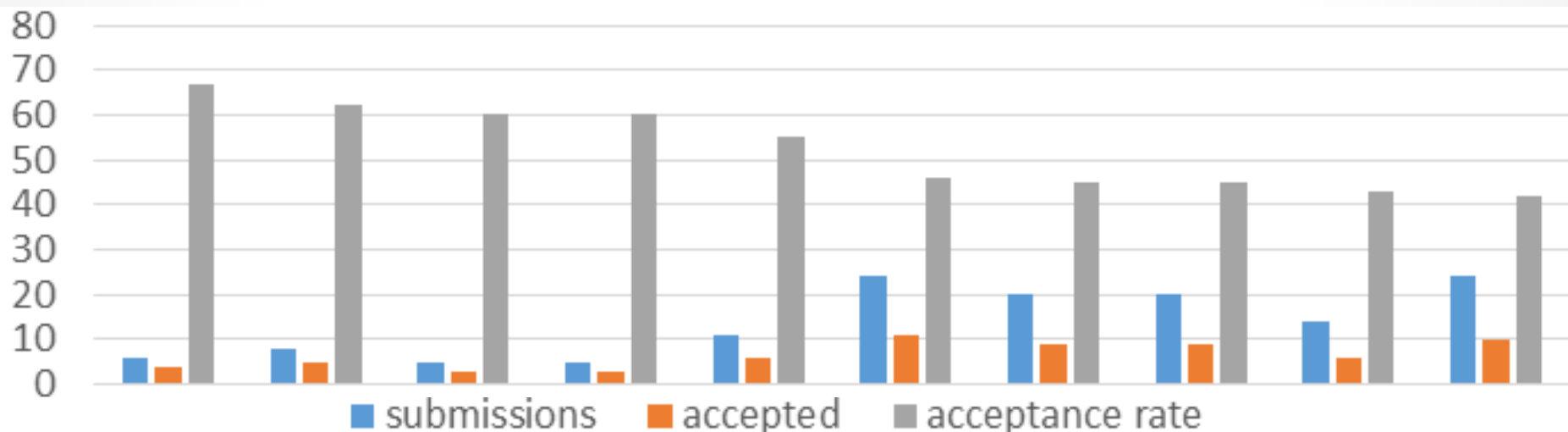
行動
分析



採択されやすいトピック

流行りに乗らないのが重要?

エンゲージメント	セッション検索	検索インタフェース	実験設計	知識グラフ	
	圧縮	知能的なパーソナライズ	ユーザスタディ	ユーザとタスクのモデル	評価尺度



採択されにくいトピック

NLP タスクが多く感じる?

情報
抽出

コミュニティ
検索

情報
要約

リンク
分析

新しい
検索タスク

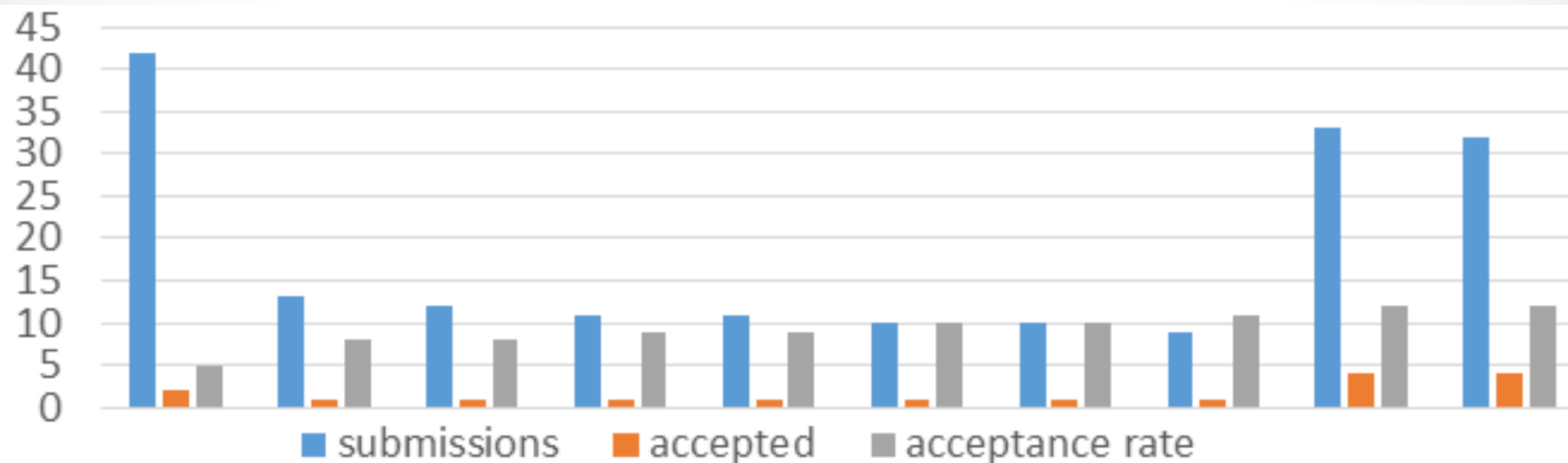
Web スケール
ランキング

図書館
情報学

意見
分析

クエリ
分析

分類



Full paper タイトル タグクラウド



頻出語除外 タグクラウド

Search, Information, Retrieval, Recommendation 除外



Best paper award

- BitFunnel: Revisiting Signatures for Search
 - Bob Goodwin, Michael Hopcroft, Dan Luu (Microsoft), Alex Clemmer (Heptio), Mihaela Curmei, Sameh Elnikety, Yuxiong He (Microsoft)
- 概要
 - 一般的にはシグネチャファイルよりも転置ファイルの方がより高性能
 - Bit-sliced signatures は高性能なシグネチャファイル方式
 - Bing search の中で使用
 - Bit-sliced signatures の制約解決やクラウド環境にも適用可能な BitFunnel 提案
 - 既存の手法よりも極めて効率的な処理が可能

SIGIR 2016 の Best paper は
fMRI を使った被験者実験

Honorable Mention 1

- IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models
 - Jun Wang (University College London), Lantao Yu, Weinan Zhang (Shanghai Jiao Tong University), Yu Gong, Yinghui Xu (Alibaba Inc.), Benyou Wang, Peng Zhang (Tianjin University), Dell Zhang (Birkbeck, University of London)
- 概要
 - IR タスクでは順位付きリスト形式で結果を提示
 - ・ テキスト検索, Web 検索, 推薦システム etc.
 - (非) 適合ラベルと文書の特徴量から教師あり学習
 - ・ ランキング学習
 - リンクやクリックログを用いて, ゲーム理論に基づき, 二種類のモデル構築手法を提案
 - ・ 識別モデル, 生成モデル
 - 評価実験の結果, いろんなタスクで精度向上

Honorable Mention 2

- Classification by Retrieval: Binarizing Data and Classifiers
 - Fumin Shen (University of Electronic Science and Technology of China), Yadong Mu (Peking University), Yang Yang (University of Electronic Science and Technology of China), Wei Liu (Tencent AI Lab), Li Liu (Malong Technologies Co., Ltd), Jingkuan Song, Heng Tao Shen (University of Electronic Science and Technology of China)
- 概要
 - 画像識別タスク
 - 画像を数千のカテゴリのうちどれかに割当
 - なぜ SIGIR に...?
 - C 個のカテゴリ・ D 次元データだと (最低) $D \times C$ 個のパラメータ
 - メモリ使用量大
 - Compact binary hash codes を使うとメモリ使用量少
 - 複数の画像のデータセットで実験した結果, 計算コスト及びメモリ使用量削減

Next SIGIR conferences

- SIGIR 2018
 - USA ミシガン州 アナーバー
- SIGIR 2019
 - フランス パリ
- SIGIR 2020
 - 中国 西安
 - SIGIR 2017 を超える参加者 (1,000 人) を目指す

Women in IR



Welcome reception





Students' get-together



Diversity & Inclusion Luncheon



Banquet



Keynote 1

- Forward to the past: notes towards a pre-history of web search
 - Stephen E. Robertson
 - BM25 (BM25F) の提案者
 - Gerard Salton Award 受賞者 (IR におけるノーベル賞)
- 概要
 - 情報検索の歴史を紹介
 - Indexing, ranking, RDB etc.
 - Google が上手く行った理由
 - クローリング技術が素晴らしかった
 - Web 検索の多様性を認識していた
 - 自然文クエリ・フレーズクエリの分析
 - アンカーテキスト (リンク) 利用
 - PageRank の (ランキングへの) 貢献はさほど大きくない
 - (後年は) ユーザのフィードバック学習

Keynote 2

- Mail Search: It's Getting Personal!
 - Yoelle Maarek
 - Vice-President of Research at Yahoo
- 概要
 - E-mail は廃れるという予測は多いが、実際はどんどん増加
 - 90% 以上は自動生成された文面
 - 受信ボックスはパーソナルストア
 - ショッピングレシビ、旅行保険、リマインダ etc.
 - 有用か不要かの判別が困難なジャンクも含む
 - Re-finding task では recall が極めて重要
 - Search
 - クエリは情報不足で役に立たない
 - Browse
 - フォルダ分けするユーザ 30%
 - フィルタを使うユーザ 10%
 - 結果提示
 - スコアが閾値を超えるメールだけをタイムスタンプでソート
 - 検索ログやメール本文を分析してスコアリングに利用

チュートリアル

1. Statistical Significance Testing in Information Retrieval: Theory and Practice
2. Candidate Selection for Large Scale Personalized Search and Recommender Systems
3. A/B Testing at Scale: Accelerating Software Innovation
4. Probabilistic Topic Models for Text Data Retrieval and Analysis
5. Neural Networks for Information Retrieval (NN4IR)
 - 一番人気, 参加者 200 名
 - 資料ダウンロードサイト: <http://nn4ir.com/>
6. Building Test Collections: An Interactive Guide for Students and Others Without Their Own Evaluation Conference Series
7. From Design to Analysis: Conducting Controlled Laboratory Experiments with Users

Building Test Collections: Ian Soboroff (NIST)

- 補足

- NIST は TREC を作成する機関
- テストコレクションの構成要素
 - ・ 文書集合
 - ・ クエリ
 - ・ 適合性評価データ

- 内容

- (参加者含) 自己紹介
- テストコレクションの作り方
- テストコレクションの妥当性検証方法
- グループアクティビティ
 - ・ 各自のタスクを共有
 - ・ 適切なテストコレクションを検討

- アセスメント方法

- 少数の人を訓練して評価もクラウドソーシングもコストがかかるのは同じ

- 印象的なお言葉

- “困ったらいつでも連絡してくれ。もし返事がなければ催促してくれていい。それが俺の仕事だ”



ワークショップ

1. ATIR'17: Axiomatic Thinking for Information Retrieval and Related Tasks
2. BIRNDL'17: Bibliometric-enhanced IR and NLP for Digital Libraries
3. CAIR'17: Conversational Approaches to Information Retrieval
4. ECOM'17: eCommerce
5. KG4IR'17: Knowledge Graphs and Semantics for Text Retrieval and Analysis
6. LIARR'17: Lucene for Information Access and Retrieval Research
7. Neu-IR'17: Neural Information Retrieval
 - 一番人気
8. OKBQA'17: Open Knowledge Base and Question Answering

参加者アンケート

- 有効回答数 184/911 (20%)
- 95% の回答者が会議に満足
- ポジティブコメント
 - 過去最高だった
 - 学生ボランティアできて光栄
 - ソーシャルイベントが多くてよかった
 - 完璧なアトラクションに感動した！台風と地震最高だったぜ
 - 食事が提供されたのと都会だったので COLING'16 よりもよかった
- ネガティブコメント
 - 料理はもっと改善できた
 - ソーシャライズの仕組みを設けて欲しかった



Supporters

- ヤフー株式会社
- Baido
- Alibaba Group
- Amazon
- Sogou
- 楽天
- Wider Planet Inc.
- Microsoft Corporation
- Yahoo Research
- Google
- IBM
- eBay
- NAVER Corportion
- Huwawei Technologies
- Yandex
- LINE Corporation
- 日立
- JAIST (Wiley Co.)
- Facebook
- Morgan & Claypool Publishers
- Springer
- NOW Publisher

Institutional supporters

- 東京都観光財団
- NICT国際交流プログラム
- KDDI財団
- 栢森情報科学振興財団
- 言語処理学会
- 国立情報学研究所

Academic supporters

- 情報処理学会
- 情報処理学会データベースシステム研究会
- 情報処理学会情報基礎とアクセス技術研究会
- **電子情報通信学会**
- 電子情報通信学会データ工学研究専門委員会
- 人工知能学会
- 日本データベース学会
- ACM SIGMOD 日本支部
- 日本図書館情報学会
- 筑波大学知的コミュニティ基盤研究センター

論文紹介 1

- Efficient Cost-Aware Cascade Ranking in Multi-Stage Retrieval
 - Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, J. Shane Culpepper (RMIT University)
- 補足
 - 情報検索において二段階の結果表示方法が存在
 - 検索対象の取得
 - 検索対象のランキング
- 概要
 - (非) 適合ラベルと特徴量からモデルを学習
 - ランキング学習
 - LambdaMART, Gradient Boosted Decision Trees
 - 複数回のランキング学習を行う
 - Efficiency 確保のため re-rank する文書の件数を調整
 - 一般的な情報検索では、上位の検索結果が重要

論文紹介 2

- A Study of Snippet Length and Informativeness
 - David Maxwell, Leif Azzopardi, Yashar Moshfeghi (University of Glasgow)
- 概要
 - スニペット: 検索結果ページの文書の要約
 - 適切なサイズを実験的に評価
 - ユーザはより長い結果を好む (クリック)
 - 情報量が多いと認識
 - 非適合文書の場合でも同様

“**文書**”検索は終わったという意見があったり
モバイル検索の増加から“**情報**”検索の必要性

論文紹介 3

- On the Power Law Frequency Distribution

- Flavio Chierichetti
 - Ravi Kumar (Google)

- 補足

- Zipf の法則

- 出現頻度が k 番目のデータは全体の $1/k$ を占める
 - 情報検索のいろいろなモデルもこの法則を仮定

- 概要

- Zipf の法則に則らないデータセットたくさん存在
 - ノイズとなる 2 種類現象が確認
 - ノイズを吸収するようなモデルを提案

