# Web N-gram Workshop 2010

**Chengxiang Zhai**
University of Illinois
at Urbana-Champaign
*czhai@cs.uiuc.edu*

**David Yarowsky**
Johns Hopkins University
*yarowsky@cs.jhu.edu*

**Evelyne Viegas**
Microsoft Research
*evelynev@microsoft.com*

**Kuansan Wang**
Microsoft Research
*kuansanw@microsoft.com*

**Stephan Vogel**
Carnegie Mellon University
*stephan.vogel@cs.cmu.edu*

**Abstract**
The Web N-gram Workshop was held on July 23, 2010 in Geneva, Switzerland, in conjunction with the 33rd Annual ACM SIGIR Conference. The workshop brought together leaders in information retrieval and language modeling to discuss the challenges in information retrieval and how language modeling approaches may help address some of these challenges, with a focus on using n-gram model to help solve them. The workshop consisted of 1 invited talk, 1 tutorial, 1 panel, 10 refereed paper presentations along with discussion sessions built-in in the agenda.

## 1 Introduction

The goal of the workshop was to bring together a group of leaders in information retrieval and language modeling to discuss the challenges in information retrieval and how language modeling approaches may help address some of these challenges. The focus was on the use of n-gram models to further research in areas such as document representation and content analysis (e.g., clustering, classification, information extraction), query analysis (e.g., query suggestion, query reformulation), retrieval models and ranking, and spelling as well as the access to n-grams as an enabler of experimental design.

Often discussed in the research community is the lack of large-scale dataset and benchmarks to run experiments. This workshop addressed this issue by bringing together the community of researchers who use n-grams, already made available by Yahoo and Google/LDC along with a new Web N-gram service through which Microsoft Research, in partnership with Microsoft Bing, is providing the research community access to petabytes of Web N-gram data via a cloud-based platform.

The Web N-gram services directly address the data need by enabling the community of researchers to create data benchmarks for repeatable experiments, and by enabling the research community to be at the forefront of inventions based on real-world, large-scale data.

Previous efforts of delivering n-grams to the research community adopted a data release approach with a cut off on the n-gram counts that obfuscate the long tail effects, an issue this service-based approach makes possible for further studies. Moreover, previous efforts also focused on just the document body; whereas richer types of textual contents are included in the Web N-gram service that can engage researchers in new innovations.

Another notable difference is the scale: the Web N-gram service provides access to petabytes of data via services—up to two orders of magnitude greater than currently available offerings. Finally, by providing regular data refresh, the Web N-gram service can open up new research directions in fields where lack of dynamic data has locked academic researchers into conducting research over static and stale data sets.

Topics addressed include  the use of the Microsoft Web N-gram services to explore novel applications of language models (e.g., long tail effects) and use of these data for enhancing the search experience (e.g., use of anchor text as a proxy to queries).

The workshop also included research and experiments using or comparing different n-grams data sets which can help create a useful n-gram baseline for the research community, in terms of n-gram attributes such as size, access, content, and model types needed for researchers.

Other topics discussed in the research papers presented at the workshop include using n-gram language model for information extraction, Web information retrieval, helping people with writing in a foreign language, and efficient storage of large language models.

# 2 Program

The program was designed to emphasize interactions between the participants. To that end, the workshop presenters were invited to present their research via long or short talks, and a panel was added as a means to engage all the workshop participants in the debate of accessing data via data release versus data services. A hands-on tutorial on the Microsoft Web N-gram Services served as a concrete example of large scale data access via live data services.
The day was kicked off by an invited talk to help set the tone of the workshop by considering the living web as the corpus of study.

## 2.1 Invited talk, tutorial, panel

### 2.1.1 Invited Talk

**Beyond Googleology: Assessing the Composition of the Web as a Large Corpus**
Serge Sharoff, Centre for Translation Studies, University of Leeds

As the first talk of the workshop program, Dr. Sharoff's invited talk served very well to motivate the overall theme of the workshop and offered many intriguing examples of leveraging the Web as a huge corpus to extract linguistic knowledge.  A main point made in the talk was that the Web can be treated as a huge repository of corpora covering multiple languages and different genre, raising the interesting challenge of how to tap into this ever-growing valuable resource effectively and efficiently. Using the current search engines to extract various patterns and statistics is convenient, but insufficient. Thus "live services" that intend to provide systematic access to fresh Web data, such as the Microsoft Web n-gram language model service, can be expected to be very useful.

### 2.1.2 Tutorial

**An Introduction to Web N-gram Service**
Kuansan Wang, Microsoft Research

The goal of the tutorial was to give an introduction to the Microsoft Web N-gram Services in terms of service description, access and the documentation available. It emphasized the data available to the community, and introduced a new service besides the Lookup service, namely: the Generative Service. Details on the data made available through the services and service description can all be found at http://research.microsoft.com/web-ngram.

### 2.1.3 Panel

**Web-based Data Services for Research – Challenges and Opportunities**
*Moderator:* Stephan Vogel, Carnegie Mellon University

*Participants:*
Kenneth Church, John Hopkins University
Evgeniy Gabrilovich, Yahoo! Research
Haym Hirsch, National Science Foundation
Donald Metzler, Information Sciences Institute University of Southern California
Kuansan Wang, Microsoft Research

The panel was introduced by Stephan Vogel. Each panelist then presented their thoughts on what the challenges and opportunities of web-based data services create. Some of the issues raised included: data sensitivities, replicability, and services versus data.

A lively discussion followed, in which a surprisingly large number of participants contributed with comments and questions to the panelists. As expected, two standpoints were visible, one from the academia community, and one from those working at companies. Researchers at universities prefer to have full access to the data; companies have to be very careful about the legal pitfalls when making data publicly available. The discussion centered on

- the issue of services versus data: i.e. having the data provides more flexibility and control, than only having access through the data via a provided service;
- the relevance of data for research in academia, which in turn can benefit companies: e.g. search queries or user click logs;
- the legal problems of releasing data which may contain personally identifiable information.

## 2.2 Long and short talks

The workshop included four long presentations and five short presentations, further described below.

### 2.2.1 Long talks

**Information Extraction from Web-Scale N-Gram Data**
Niket Tandon, Gerard De Melo, Max Planck Institute

**A Comparative Study of Bing Web N-gram Language Models for Web Search and Natural Language Processing**
Jianfeng Gao, Patrick Ngyuen, Microsoft Research; Xiaolong Li, Microsoft Bing; Chris Thrasher, Mu Li, Kuansan Wang, Microsoft Research

**Verifying the Implicit Presence of Difficult Query Aspects using a Large External Corpus**
Dmitri Roussinov, University of Strathclyde

**Minimal Perfect Hash Rank: Efficient Storage of Large Language Models**
David Guthrie, Mark Hepple, University of Sheffield

### 2.2.2 Short talks

**Global Statistics in Proximity Weighting Models**
Craig Macdonald, Iadh Ounis, University of Glasgow

**Further Studies on Multi-Style Language Model for Web Information Retrieval**
Xiaolong Li, Microsoft Bing; Jianfeng Gao, Kuansan Wang, Microsoft Research

**Using Web N-Grams to Help Second-Language Speakers**
Martin Potthastm, Martin Trenkmann, Benno Stein, Bauhaus-Universität Weimar

**Comparing Web N-grams and Other Means of Identifying Named Entities in Corporate Blogs**
Aditya Rachakonda, Srinath Srinivasa, IIITB; Sudarshan Murthy, Wipro Technologies; Avinashreddy Palleti, Ramya Krishna, IIITB

**Language Differences and Metadata Features on Twitter**
Emre Kiciman, Microsoft Research

# 3 Committees

## 3.1 Organizing committee

Chengxiang Zhai, University of Illinois at Urbana-Champaign
David Yarowsky, Johns Hopkins University
Evelyne Viegas, Microsoft Research
Kuansan Wang, Microsoft Research
Stephan Vogel, Carnegie Mellon University

## 3.2 Program committee

Alistair Moffat, University of Melbourne
Amanda Spink, Loughborough University
Bill Dolan, Microsoft Research
Brian Davison, Lehigh University
Bruce Croft, University of Massachusetts Amherst
Charlie Clarke, University of Waterloo
ChengXiang Zhai, University of Illinois at Urbana-Champaign
David Yarowsky, John Hopkins University
Efthimis N. Efthimiadis, University of Washington
Emmanuel Prochasson, Hong Kong University of Science & Technology
Eugene Agichtein, Emory University
Evelyne Viegas, Microsoft Research
Eytan Adar, University of Michigan
Georges Dupret, Yahoo! Research
Hongyuan Zha, Georgia Tech University
Jaime Callan, Carnegie Mellon University
Jian-Tao Sun, Microsoft Research Asia
Jurgen Van Gael, University of Cambridge

Ken Church, Johns Hopkins University
Kevin Chang, University of Illinois at Urbana Champaign
Kuansan Wang, Microsoft Research
Michael Gamon, Microsoft Research
Nick Craswell, Microsoft
Peng Xu, Google Research
Stefan Vogel, Carnegie Mellon University
Thorsten Brants, Google Research

## 4 What's next based on participants' feedback

The Web N-gram SIGIR 2010 workshop helped in confirming the need for services such as the cloud-based Web N-gram services while identifying new needs for the research community.

As a result, Microsoft has made available new content (e.g. Query Language Models) and new services (e.g. Predictive API), available from http://research.microsoft.com/web-ngram.
Moreover, the organizers are planning a follow up workshop at the next SIGIR 2011.

## 5 Acknowledgements

The organizing committee would like to thank all of the speakers and participants as well as everyone who submitted a paper to the Web N-gram workshop. Special thanks to the reviewers for helping build an exciting workshop.

Last but not least, we would like to thank the SIGIR 2010 workshop co-chairs, Omar Alonso and Giambattista Amati as well as the local organizers Stephane Marchand-Maillet and Fabio Crestani for their help and guidance in putting the workshop together.