NAACL-HLT 2012

WLM 2012:
Will We Ever Really Replace the N-gram Model?
On the Future of Language Modeling for HLT

Workshop Notes

June 8, 2012
Montréal, Canada

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53707
USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN13: 978-1-937284-20-6
ISBN10: 1-937284-20-4

Introduction

Welcome to the NAACL-HLT workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT!

Language models are a critical component in many speech and natural language processing technologies, such as speech recognition and understanding, voice search, conversational interaction and machine translation. Over the last few decades, several advanced language modeling ideas have been proposed. Some of these approaches have focused on incorporating linguistic information such as syntax and semantics whereas others have focused on fundamental modeling and parameter estimation techniques. Although tremendous progress has been made in language modeling, n-grams are still very much the state-of-the-art due to the simplicity of the model and good performance they can achieve.

The aim of this workshop is to bring together researchers from natural language processing, linguistics and spoken language processing and to provide a venue to explore and discuss new approaches to language modeling for different applications. We have received an excellent set of papers focused on neural network language models, discriminative language models and language models using explicit syntactic information. Some of the papers investigated these models with different architectures, while others used large scale and unsupervised set-ups.

Our workshop will feature two keynote talks. We begin with a keynote from Shankar Kumar (Google Inc.) and will conclude with the second keynote from Brian Roark (Oregon Health and Science University). We will close with an open discussion led by several prominent researchers that will summarize the emerging areas of research in language modeling, the issues and challenges for various tasks that we have learnt/highlighted over the course of this workshop, common resources and evaluation challenges that can be posed to the research community.

With the advent of smart phone technologies and increased use of natural language interactions for many day-to-day tasks, it is the correct time to bring together experts in the fields of linguistics, speech and natural language processing and machine learning from industry and academia to share their ideas and set the stage for the future of language modeling.

We are especially grateful to the program committee for their hard work and the presenters for their excellent papers.

Organizing Committee

Bhuvana Ramabhadran, Sanjeev Khudanpur and Ebru Arisoy

Organizers:

Bhuvana Ramabhadran, IBM T.J. Watson Research Center (USA)
Sanjeev Khudanpur, Johns Hopkins University (USA)
Ebru Arisoy, IBM T.J. Watson Research Center (USA)

Program Committee:

Ciprian Chelba, Google Inc. (USA)
Stanley F. Chen, IBM T.J. Watson Research Center (USA)
Ahmad Emami, IBM T.J. Watson Research Center (USA)
Tomas Mikolov, Brno University of Technology (USA)
Hermann Ney, RWTH Aachen University (Germany)
Patrick Nguyen, Google Inc. (USA)
Kemal Oflazer, Carnegie Mellon University (Qatar)
Brian Roark, Oregon Health and Science University (USA)
Murat Saraclar, Bogazici University (Turkey)
Holger Schwenk, University of LIUM (France)
Peng Xu, Google Inc. (USA)
Geoffrey Zweig, Microsoft (USA)

Invited Speakers:

Shankar Kumar, Google Inc. (USA)
Brian Roark, Oregon Health and Science University (USA)

Table of Contents

Workshop Program

Friday, June 8, 2012

9:15-9:30 Opening Remarks

9:30-10:30 Invited Talk

10:30-11:00 Coffee Break

\+ Morning Session

11:00–11:25 *Measuring the Influence of Long Range Dependencies with Neural Network Language Models*
Hai-Son Le, Alexandre Allauzen and François Yvon

11:25–11:50 *Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation*
Holger Schwenk, Anthony Rousseau and Mohammed Attik

11:50–12:15 *Deep Neural Network Language Models*
Ebru Arisoy, Tara N. Sainath, Brian Kingsbury and Bhuvana Ramabhadran

12:15-14:00 Lunch

\+ Afternoon Session

14:00–14:25 *A Challenge Set for Advancing Language Modeling*
Geoffrey Zweig and Chris J.C. Burges

14:25–14:50 *Unsupervised Vocabulary Adaptation for Morph-based Language Models*
André Mansikkaniemi and Mikko Kurimo

14:50–15:15 *Large-scale discriminative language model reranking for voice-search*
Preethi Jyothi, Leif Johnson, Ciprian Chelba and Brian Strope

15:15–15:40 *Revisiting the Case for Explicit Syntactic Information in Language Models*
Ariya Rastrow, Sanjeev Khudanpur and Mark Dredze

15:40-16:00 Coffee Break

Friday, June 8, 2012 (continued)

16:00-17:00 Invited Talk

17:00-18:00 Closing Remarks