Abstract:
We present a class of latent embedding models that are discriminatively trained to map from the content in a query-document or document-document pair to a ranking score. Like latent semantic indexing (LSI), our models take account of correlations between words (synonymy, polysemy). However unlike LSI, our models are trained with a supervised signal directly on the task of interest, which we argue is the reason for our superior results. We provide an empirical study on several tasks, including document retrieval, image annotation and ranking and music recommendation.

We also describe several extensions:
- Optimizing the top of the ranked list.
- The nonlinear case.
- Dealing with ambiguity in the query.
- Providing diversity in the results.

This is joint work with Bing Bai, David Grangier, Ronan Collobert, Kunihioko Sadamasa, Yanjun Qi, Corinna Cortes, Mehryar Mohri, Samy Bengio, Aurelien Lucchi, John Blitzer and Nicolas Usunier.

Bio:
Jason Weston is a Research Scientist at Google NY since July 2009. He earned his PhD in machine learning at Royal Holloway, University of London and at AT&T Research in Red Bank, NJ (advisor: Vladimir Vapnik) in 2000. From 2000 to 2002, he was a Researcher at Biowulf technologies, New York. From 2002 to 2003 he was a Research Scientist at the Max Planck Institute for Biological Cybernetics, Tuebingen, Germany. From 2003 to June 2009 he was a Research Staff Member at NEC Labs America, Princeton. His interests lie in statistical machine learning and its application to text, audio and images. Jason has published over 80 papers, including best paper awards at ICML and ECML.