



Spring 2009 CIS Colloquium Series

Modeling Text Quality in Newspaper Text and Machine Translation

Ani Nenkova
(University of Pennsylvania)

March 18, 2009

Abstract: The goal of our work is to develop models of text quality, including coherence, fluency and general readability. Models of text quality are a critical component for a range of text producing applications such as summarization, machine translation and text generation.

For newspaper text, we combine lexical, syntactic, and discourse features to produce a highly predictive model of human readers' judgments of text readability. This is the first study to take into account such a variety of linguistic factors and the first to empirically demonstrate that discourse relations are strongly associated with the perceived quality of text. We show that various surface metrics generally expected to be related to readability are not very good predictors of readability judgments in our Wall Street Journal corpus. We also establish that readability predictors behave differently depending on the task: predicting text readability or pairwise comparison of readability. Our experiments indicate that discourse relations are the one class of features that exhibits robustness across these two tasks.

In the context of machine translation, we study sentence fluency, which is an important component of overall text readability. We report the results of an initial study into the predictive power of surface syntactic statistics and language model features to predict fluency originally assessed for the purpose of evaluating machine translation. We find that these features are weakly but significantly correlated with readability. Machine and human translation can be distinguished with accuracy over 80% and performance on pairwise comparison of fluency is also very high, over 90%.

Bio: Ani Nenkova is assistant professor at the Department of Computer and Information Science at the University of Pennsylvania. She obtained her PhD degree from Columbia University where she worked on text summarization (evaluation, content selection and automatic rewrite). Before joining Penn she spent a year and a half at Stanford University, working on predicting and detecting prosodic prominence.

Location: 4th Floor Conference Room (Wachman 447)

Time: 12-1pm, Wednesday, March 18, 2009

Refreshments will be served!