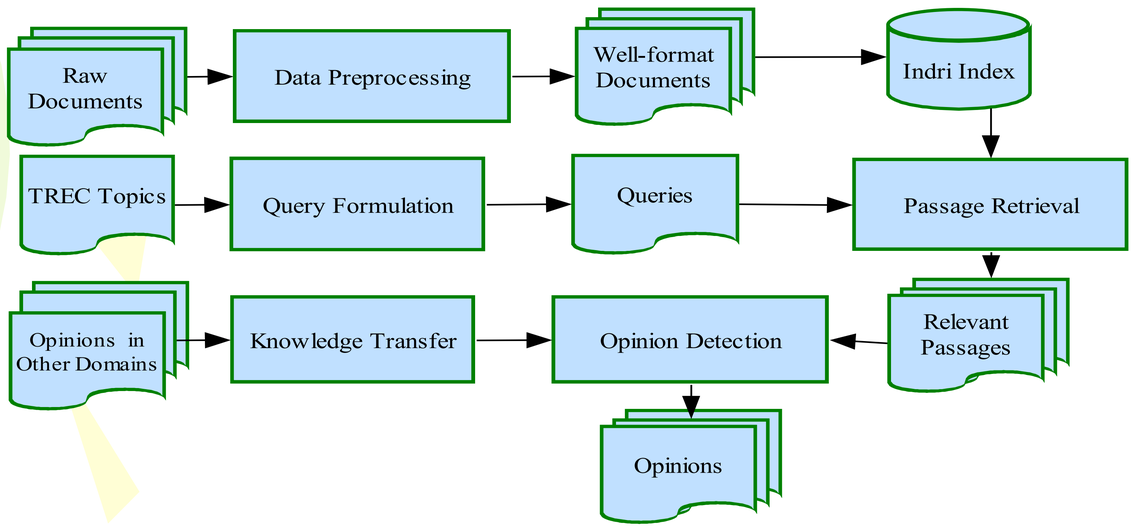
Knowledge Transfer and Opinion Detection in Blog Track

Hui Yang
Language Technologies Inst.
School of Computer Science
Carnegie Mellon University
huiyang@cs.cmu.edu

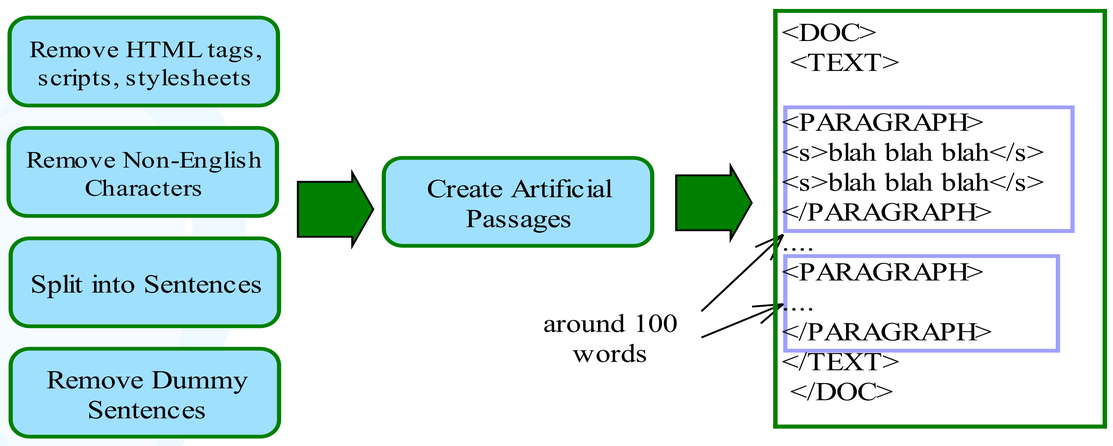
Luo Si
Dept. of Computer Science
Purdue University
lsi@cs.purdue.edu

Jamie Callan
Language Technologies Inst.
School of Computer Science
Carnegie Mellon University
jamie@cs.cmu.edu

Overview



Data Preprocessing



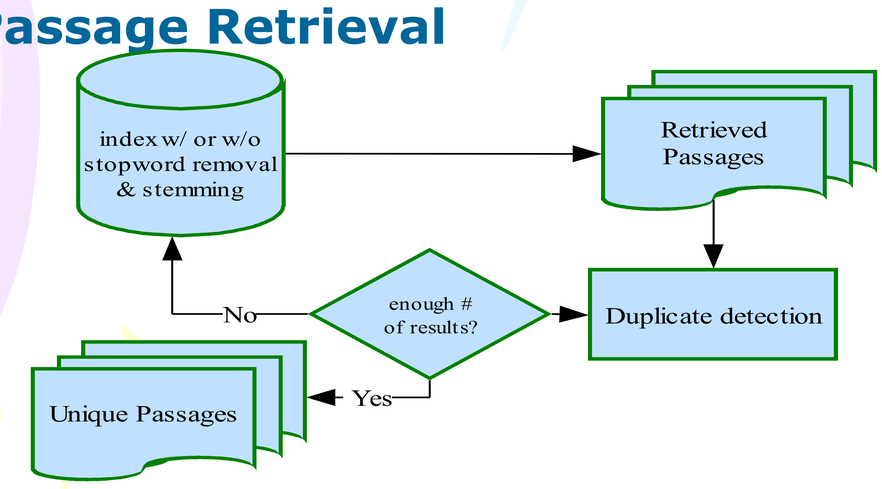
Query Formulation

- title query; e.g., *#combine[paragraph](#3(march of the penguins))*
- title query + nouns, adjectives in description; e.g., *#combine[paragraph](#3(march of the penguins) documentary film)*
- title query + opinion words; e.g., *#combine[paragraph](#uw15(#3(march of the penguins) love) #uw15(#3(march of the penguins) like) …)*
- title query + nouns, adjectives in descriptions + opinion words. e.g., *#combine[paragraph](… #uw15(#3(march of the penguins) great) #uw15(#3(march of the penguins) awesome) documentary film)*

Opinion Word Dictionary

#positive verb	#negative verb	#positive adjective	#negative adjective
Love, like	Hate, dislike	Good, best, better, happy, extraordinary, successful, glad, desirable, worthy, remarkable, funny, lovely, entertaining, decent, beautiful, fascinating, brilliant, gorgeous, perfect, nice, fantastic, impressive, fabulous, amazing, desirable, excellent, great, awesome, splendid, distinctive	Bad, awful, suck, worse, worst, poor, annoying, stupid

Passage Retrieval



Knowledge Transfer and Opinion Detection

•Training data from Movie Review

-Source: IMDB movie database
http://www.cs.cornell.edu/People/pabo/movie-review-data/
5000 opinion sentences, *5000* objective sentences

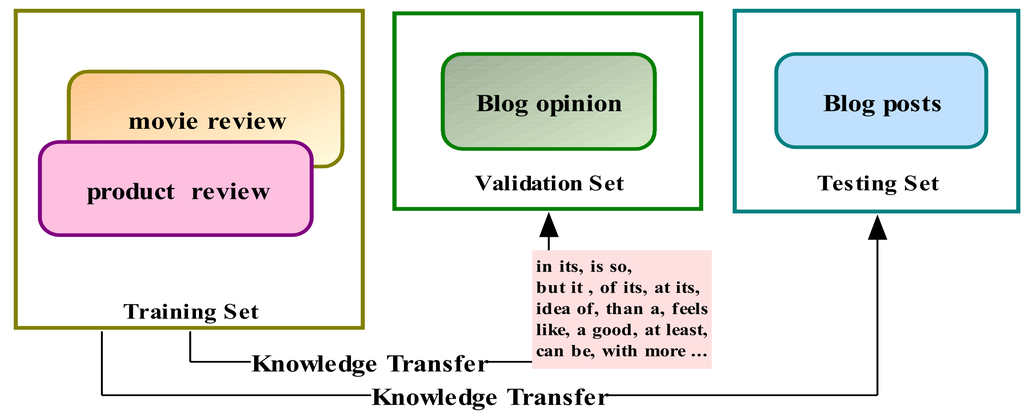
•Training data from Product Review

-Source: Amazon.com, (5 electronic products including 2 digital cameras, 1 cell phone, 1 DVD players and 1 jukeboxes)
- http://www.cs.uic.edu/~liub/FBS/FBS.html
- *2041* opinion sentences, *2217* objective sentences

•Validation data from Blog Opinions

-Source: Opinions manually selected for 11 own topics: *"Steelers", "Christmas", "Mr. and Mrs. Smith", "Harry Potter", "Condoleezza Rice", "Canon camera", "China rise", "Harvard University", "Hash brown", "Seattle",* and *"Zoloft"*
- *1201* opinion sentences, *1240* objective sentences

•Knowledge Transfer by Gathering Common N-gram Features across Different Domains



•Rank Sentences by Bayesian Logistic Regression

$$L = \Pr(Y = 1 \mid S = s, \theta) = \frac{1}{1 + \exp(-\theta^T s)}$$

$$\theta_{\max} = \arg\max_{\theta}\left(\sum \log \Pr(Y = y \mid S = s, \theta) - \frac{1}{2\sigma^2}\left\|\theta\right\|_2^2\right)$$

Y=1 when sentence S is an opinion; Y= -1 when S is an objective sentence

•Average Sentence scores to get Passage Score

Results

•The MAPs of all submitted runs are better than median
•Our best run is *r2*, which uses bigram model + adjective percentage w/o query expansion
•Opinion detection is effective, low recall of it is due to low recall of topic retrieval