UMass at TDT 2000

James Allan and Victor Lavrenko
(with David Frey and Vikas Khandelwal)

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst

Work on Story Link Detection

- Active work on SLD
  - Not ready in time for official submission
- Story "smoothing" using query expansion
- Score normalization based on language pair

What is LCA?

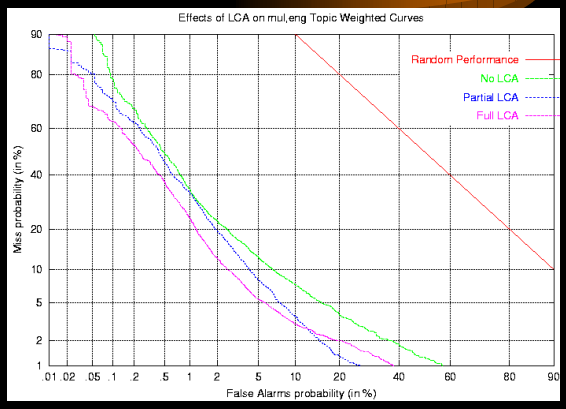
- Local Context Analysis
  - Query expansion technique from IR
  - More stable than other "pseudo RF" approaches
  - Application for more than document retrieval
- Basic idea
  - Retrieve a set of passages similar to query
  - Mine those passages for words near query
    - Ad-hoc weighting designed to do that
  - Add words to query and re-run

LCA for story smoothing

- Convert story to a weighted vector
  - Inquery weights (incl. Okapi tf component)
- Select top 100 most highly weighted terms
- Find top 20 stories most similar (cosine)
- Weight all terms in top 20 stories (LCA)
- Select top 100 LCA expansion terms
- Add to story (decaying weights from 1.0)
- Story now represented by 100-200 terms
- Compare smoothed story vectors

Smoothing SLD with LCA

- Run on training data (english)
- Green line is no smoothing
- Blue is smooth with past stories
- Pink is smooth with whole corpus (cheating)



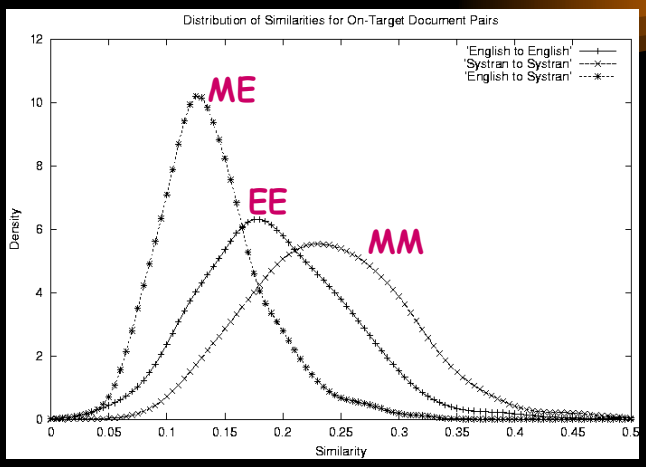
Work on Story Link Detection

- *Story "smoothing" using query expansion*
- Score normalization based on language pair

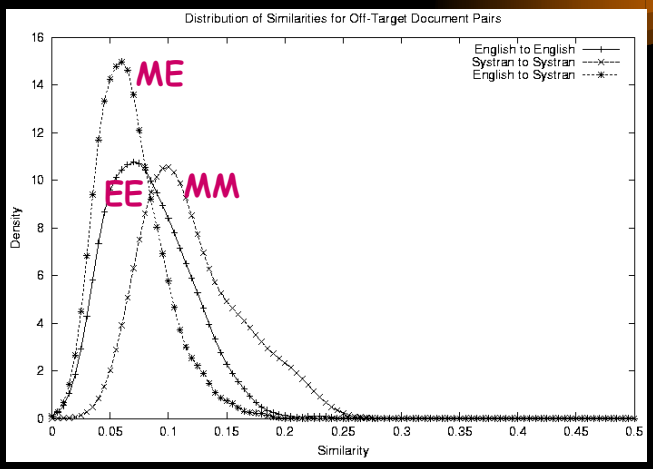
Score normalization

- Noticed that SYSTRAN documents were throwing scores off substantially
  - Multilingual SLD was much worse that ENG only
- Look at distribution of scores in same-topic and different-topic pairs

Score distributions, same topic



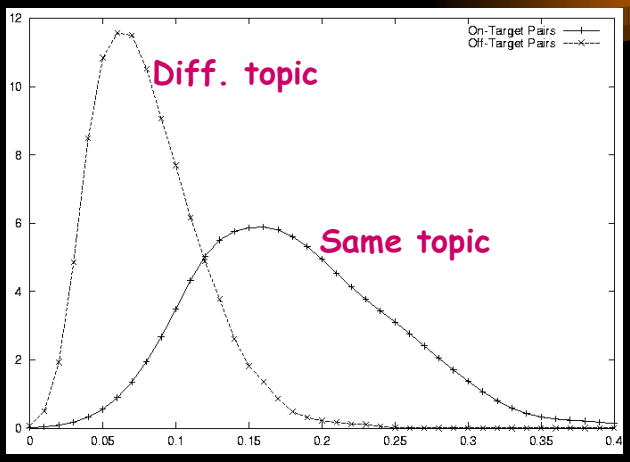
Score distributions, diff topic



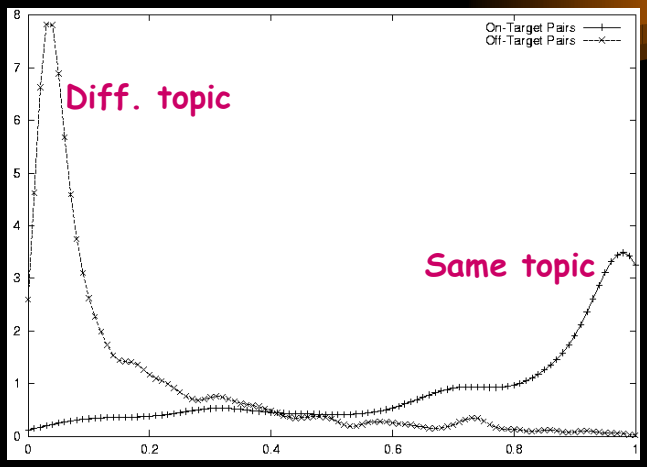
Clearly need to normalize

- SYSTRAN stories use different vocabulary
  - Stories are much more likely to be alike
  - And much less likely to be like true English
- Develop normalization based on whether within or cross-language
- Convert scores into probabilities
  - Use distribution plots for each case

Combined distribution (before normalization)



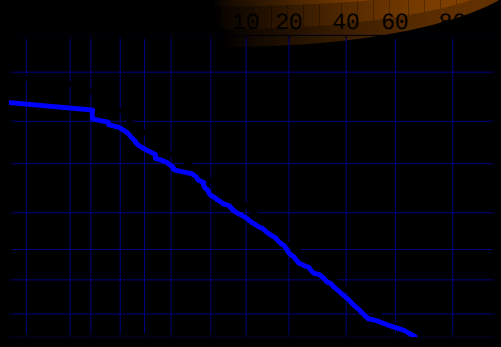
After normalization (on same data--"cheating")



Probabilities!

DET plots from normalization

- Huge change in distributions
- Less pronounced change in DET plot



Conclusions

- Story smoothing with LCA works
  - Need to "smooth" with all stories before *later*
  - Need to use different matching for smoothing and then story-story comparison
- Score normalization has potential
  - Other sites have found similar effects
  - Experiments on source-type (audio, newswire) within language pairs have been inconclusive
    - Not much training data for doing conversion