TDT 2000 Workshop Lessons Learned

These slides represent some of the ideas that were tried for TDT 2000, some conclusions that were reached about techniques on some tasks, and various other thoughts on the tasks. In general, the items here arose during presentations or during discussions following each task. These represent the impressions of the group (though mostly of the person typing: me) and their accuracy may not be perfect. Please take them in that spirit.

James Allan, November 2000

Goals of meeting

- Discuss TDT 2000 evaluation
- Decide on any lessons learned
  - Potential for HLT conference?
- Relate TDT to TREC filtering
  - Including discussions of merging
- Decide on TDT 2001 evaluation
  - Reality: little to no funding for new data
- Look ahead to TDT 2002 (?!?)

Corpus

- Impact/quality of search-guided annotation?
- New TDT-3 topics substantially different in quality from old 60
  - Different numbers of stories in E+M
- TDT-2 as training/dev data
  - May/June has only 34 topics, AMJ has 69

MEI (Hopkins WS'00)

- Strictly cross-language tracking (E→M)
  - Point: varying N_t stories is like query track
- Phrase translations by dictionary inversion
- Results
  - Phrases beat words
  - Translation preferences (for effectiveness)
    - Phrases then words then lemmas then syllables
  - Post-translation re-segmentation
    - Char bigrams are best, syllable bigrams do poorly

Tracking—what people did

- Models
  - Vector space
  - Clusters, Okapi-esque weights
  - Statistical language model
    - Likelihood, story length, score normalization (all of 'em)
- Use detection system—N_t seed cluster(s)
- Cluster on-topic stories (v. 1-NN)
  - Advantage to merging new stories into topic, but heavily weighting N_t stories
- Putting N_t stories into variable clusters

Tracking

- Named entities as features
  - Helped when added to morphs+stemming (IBM)
  - High miss rate when only NE's used: many stories have no NE's in common (Iowa)
    - Better for newswire in English
- Use of χ^2 for query term selection
  - Used negative exemplars to improve

Tracking (cont)

- Negative exemplars helped for English (UMd)
  - Not for Mandarin, but perhaps too much noise
- Character bigrams much better than words
- Improvements in translation help performance
  - Particularly at the lower miss rates

Tracking lessons

- Pretty much matched TDT 99 results
  - In sense of getting into 10m/1fa box
  - With N_t=1 (this year) vs. N_t=4 (then)
- Automatic story boundaries have noticeable impact on effectiveness
  - Not huge, but ref boundaries dominate
  - Not as clear for English (tuning issue only?)
- Variability of BBN's system with different N_t stories selected
  - Suggests variability based on sample stories
    - Should have have various samples for running?
    - A way to get zillions of "other" topics to track

Tracking (cont.)

- Stemming helped (on TDT-2)
- Challenge (N_t=4, ASR, no boundaries) sometimes better or no worse than primary condition (N_t=1, CCAP, ref boundaries)
- NE's contain useful info, but not enough
- Negative exemplars may help
- Translation matters (only at low miss rates?)
- Score normalization continues to be an issue

Tracking questions

- Impact of topic size on effectiveness
  - Evidence that small topics easier
- "Value" of score normalization
  - Per-topic "dial" vs. per-system "dial"

First Story Detection

- UMass improved slightly
- ASR hurts slightly
- Automatic boundaries hurt slightly

Cluster detection—recap/summary

- Sentence boundaries (CUHK)
- Named entities (English and Mandarin)
  - Learned on training corpora (CUHK)
- Translation (M→E)
  - Dictionary, also parallel corpus (passage-aligned)
    - Used to adjust weights of dictionary-translated words
      - Seems to help (though baseline cost is high)
- Use of deferral window (temporary clusters)
  - Seems reasonable, but value unclear

Cluster detection (cont.)

- Interpolation rather than backoff
  - Backoff = get missing terms' stats from GE
  - Interpolation = *all* scores comb of cl & GE
- "Targeting" (cf. "blind RE")
  - Smooth incoming story with info from another corpus (15% from there is best)
- 20% degradation due to [these] automatic boundaries
  - Stemming hurts for auto boundaries
    - Stemming is a recall-enhancing device, so P(fa) higher

Cluster detection

- Cost increases when using native orthography
  - SYSTRAN makes a big difference
- Bigger topics tend to have higher costs
  - Easier to split a big topic?
  - Huge cost of a miss?
- 1-NN non-agglomerative approaches are not stable
  - Hurt by automatic boundaries in particular

Cluster detection

- Hurt by including Mandarin docs with English
- Hard to compare clustering by subsets
  - I.e., cannot figure out effectiveness on X by extracting those results from X+Y results
    - Including Y into a cluster impacts following X's

Cluster detection questions

- ~~(For George)~~ Real task is multi-lingual
  - SYSTRAN is just a method to get there
    - Despite Jon's breaking it out separately
    - Really a contrastive run
- Measuring effectiveness
  - Cost seems "bouncy", YDZ of unclear value
  - Minimum cost includes (say) 633 *and* 2204
    - Small changes in C_{det} ← *huge* change in #clusters
  - TREC filtering's utility measures similarly unstable
  - Oasis experience (UMass)
  - Need "better" application model?

Segmentation

- Fine-grained HMM
  - Model position in story
    - 250 states for start, 0 for end, 1 for middle
      - End states become events occurring later (at start)
  - Model where-in-story-we-are features
  - Single coherent segmentation of text
- Visualization tools
- No use of audio information (except X)

Link Detection

- Lack of interest—why?
- UMass
  - *Much* better on E-E, than M-M or M-E
  - Normalization as f(EE,MM,ME) is important
  - LCA smoothing ("targeting") helpful
    - Issue: how to find smoothing stories vs. how to compare smoothed stories

Event granularity

- Some events (e.g., Pinochet) seem to have several clear sub-topics over time
  - Clear representation of topic evolution?
- Others are much more scattered (e.g., Swiss Air crash)
- http://www.ldc.upenn.edu/Projects/Topic_Gran/
  - Currently password protected