Link Detection – Results and Analysis

Ralf D. Brown, Thomas Pierce, Yiming Yang, Jaime G. Carbonell

Language Technologies Institute
Carnegie Mellon University (CMU)
Pittsburgh, PA 15213, USA
{ralf,tomp,yiming,jgc}@cs.cmu.edu

ABSTRACT

This paper describes the two Story Link Detection systems Carnegie Mellon University (CMU) developed, and examines why their performance on the evaluation data was considerably worse than expected while performance on an alternate evaluation set matched the performance on the training data.

1. Introduction

Carnegie Mellon University submitted two systems to the official TDT-3 evaluation for the Story Link Detection (SLD) task. The two systems were independent implementations of essentially the same method with differing attempted enhancements. Development of multiple systems was enabled by a common code library for loading TDT story collections and processing the test data file; the only code which must be implemented for each new SLD system is the actual similarity/confidence computation. This common library is an outgrowth of the DTREE topic tracker from the TDT-2 project [1, 4].

These systems were run on three distinct data sets. The first ("dry run") consisted of story pairs selected from the six months of news stories originally collected for the TDT-2 project in 1998; the second ("evaluation") consisted of previously-unseen pairs selected from an additional three months of previously unseen news stories; and the third ("alternate") consisted of additional pairs selected from the same three months of new data, which was provided in response to the dismal performance on the evaluation set of all submitted systems. Both the dry run and alternate sets selected their story pairs from among those stories which had received event labels for the tracking task, while the evaluation set contained 120 candidate matches for each of 180 seed stories selected at random.

2. System Descriptions

Both of the systems submitted for the TDT evaluation used the common library to load the story collections and additional common code to provide those portions of the test and control mechanism for the story link detection task which are independent of the actual similarity determination. Because of this common code, the two systems also have many of the same capabilities, such as stop-wording and table-based stemming of the stories as they are loaded, and the ability to automatically select the optimum thresholds for declaring a pair of stories to be linked.

Note the use of the word "thresholds" in the previous sentence. The decision threshold is actually a split threshold, with different values depending on whether or not the two stories come from the same source; this permits a laxer threshold when the two stories are from sources which may have different styles. For the purposes of selecting which threshold to use, multiple TDT sources may be treated as a single source; training on the dry run data, it was determined that performance of our first system was best (for the default deferral of ten source files) if the New York Times and AP newswire were treated as one source and *all other* TDT sources combined into a second source.

The first of our systems, identified as CMU-1 in the evaluation and using the system identifier COSINE, uses incremental TF*IDF-weighted cosine similarity measures to determine whether or not two documents discuss the same topic. The stop-worded news stories are converted into binary term vectors (any nonzero number of occurrences reduced to 1) which are then weighted by the TF*IDF value of each term. To decide whether two stories are linked, the cosine similarity measure – the normalized inner (dot) product – of the two corresponding term vectors is computed, and a YES decision output if the similarity is above a predetermined threshold.

For the evaluation, the TF*IDF values for COSINE were initialized from the complete collection of English documents in the dry-run data set which was available for training. As the evaluation data was processed, the TF*IDF values were incrementally updated to adapt to changing patterns of use over time. The COSINE system can additionally apply a time-based decay to the similarity score, making temporally distant story pairs less likely to be declared linked, but this feature was not used as it was determined to be detrimental to the cost measure during early testing.

The second system, identified as CMU-2 in the evaluation, is also based on weighted cosine similarity measures, though with different weighting and thresholds. Unlike the COSINE system, the logarithm of the term frequency was used ('ltc' rather than 'ntc' in the terminology of the SMART document-retrieval system[3]), and the TF*IDF statistics were derived solely from the test stories as they were processed, rather than having been initialized from the six-month training corpus. The CMU-2 system additionally contains a probabilistic modeler, which was disabled for the evaluation.

3. Performance

Table 1 lists the results of the seven runs CMU submitted to the December 1999 evaluation, showing the normalized cost measure C_{link} for each run. While the CMU-2 system performed worse than it had on the training data, it still did much better than CMU-1, which had been designated CMU's official system for the evaluation. In fact, CMU-1 had a cost measure worse than the strawman strategy of *never* indicating that stories are linked ("Just Say No"). The majority of this difference in performance is due to the different decision thresholds selected for the two systems – CMU-2 used much higher thresholds, which proved to be quite close to the op-

System	Transcription	Deferral	Norm(C_{link})
CMU-1	ASR	1	1.1260
CMU-1	ASR	10	1.0943
CMU-1	ASR	100	1.0921
CMU-1	manual	1	1.1477
CMU-1	manual	10	1.1657
CMU-1	manual	100	1.0974
CMU-2	ASR	10	0.4667

Table 1: Official Evaluation Results

timum for the evaluation data. Tuning the CMU-1 system on the evaluation data yields a normalized C_{link} near 0.58, almost exactly a factor of two improvement.

The current (post-evaluation) best performance of the CMU-1 system with a ten-file deferral period produces a normalized C_{link} of 0.1399, 1.1320, and 0.1392 for the dry-run, evaluation, and alternate data sets when tuned on the dry-run data set. Even though the performance on the training data is now slightly better than it had been at the time of the official evaluation, performance on the evaluation data is worse because the selected thresholds have shifted even further away from the optimum for the evaluation data. The CMU-2 system currently achieves cost values of 0.1267, 1.2867, and 0.1269 for the dry-run, evaluation, and alternate data sets.

Figure 1 shows how errors and false alarms made by the CMU-1 system may be traded against one another by varying the threshold on the similarity measure for each of the three test sets. Figure 2 plots the equivalent Detection-Error Tradeoff (DET) curves for the CMU-2 system. It is obvious that the December 1999 evaluation data yields a DET curve which differs markedly from those for the other two data sets on both SLD systems; this will be examined further in the next section.

Another way to present the performance is with the F_1 measure which is commonly used in the broader information-retrieval community. F_1 is defined as $2pr/(p+r)$, where p is precision (proportion of retrieved documents which should have been retrieved) and r is recall (proportion of documents which should have been retrieved that actually were retrieved). When tuned for F_1 on the dry-run data set, CMU-1 currently achieves micro-averaged F_1 values of 0.92, 0.56, and 0.93 on the dry-run, evaluation, and alternate data sets – very good except on the evaluation set. Even when tuned on the evaluation data, F_1 for the evaluation set is only 0.70.

4. What Went Wrong?

Cross-validation on the training data using the COSINE system led us to expect that the cost measure would be 20 to 30% higher on the evaluation data than on the training data, yet the CMU-1 system had a cost some six times as high and the CMU-2 system had a cost measure more than twice as high. Why did the two systems fare so much more poorly on the evaluation data than on the training data? The simple answer seems to be that the evaluation data is dramatically different from (and "harder" than) the training data.

The dry-run training data (as well as the alternate test set created after the December 1999 evaluation) was generated by using the event-labeled stories in the English portion of the collection and associating 120 random other labeled stories with one story from each event. Of the randomly-selected stories, those which had the same label as the initial story were considered linked, while those which did not were considered not linked.

In contrast, the evaluation data was not limited to the subset of the collection which had been labeled. However, it is not clear *a priori* whether restricting the test to labeled documents helps or hurts performance. Using only labeled stories may make the decision easier, since many confounding stories would not be included in the data set, i.e. bombings other than those for which labels have been assigned. On the other hand, similar labeled events which should not be considered linked will make up a larger proportion of the reduced data set, increasing their impact.

Figure 3 compares the distribution of similarity scores computed by the CMU-1 system on each of the three data sets. The score for each story pair was placed into one of 1000 bins by truncating the score to 3 decimal places, and the number of elements in each bin was then plotted. The Y axis of each plot has been truncated somewhat to better illustrate the behavior in the region 0.1 to 0.2. As is clear from the figure, the distribution of scores are very similar between the dry run (top) and alternate (bottom) data, and quite different for the evaluation data (middle). Note the local minimum near 0.1, followed by a local maximum at 0.2 in the top and bottom graphs; this may be an indication of two well-separated Gaussian distributions for linked and non-linked story pairs.

Further indications that the evaluation data is qualitatively different from the training data are given by the dramatically higher optimal thresholds (0.18-0.22 versus 0.065-0.075 for CMU-1) and the differing effects on C_{link} of various parameter settings for those data sets. Thus, disabling TF*IDF weighting in the CMU-1 system substantially (40% or more) increases C_{link} for the dry run and alternate sets, yet moderately *decreases* the optimum C_{link} for the evaluation set, from 0.58 to about 0.50. Similarly, applying a time-based decay to the similarity score worsens C_{link} on the dry run data, but improves performance on the evaluation data – even when the system is first tuned on the very data on which it is to be tested. The latter effect indicates that the events in the evaluation data are much more temporally focused than the events in either of the other sets.

5. Conclusions and Future Work

Although both Carnegie Mellon SLD systems performed very well on the training and alternate evaluation data sets, performance on the official December 1999 evaluation data leaves much to be desired. In light of these results, it is clearly imperative to find a similarity measure which is less affected by differences in the data sets.

The common code for processing the test file already supports the use of multiple decision strategies and a variety of methods for combining their outputs into a single decision (majority vote, weighted votes, all-but-one, etc.). By coding additional similarity measures which make independent errors, such a multi-strategy SLD system promises better overall performance than any one of its component strategies. Combining independent decisions has proven to be beneficial in speech recognition (the ROVER system[2]) and in CMU's own investigations on improving performance on the tracking task (described elsewhere in this volume).

An obvious extension to the existing split threshold would be to use a different threshold for each possible combination of news sources (those "sources" possibly encompassing multiple TDT sources). It is

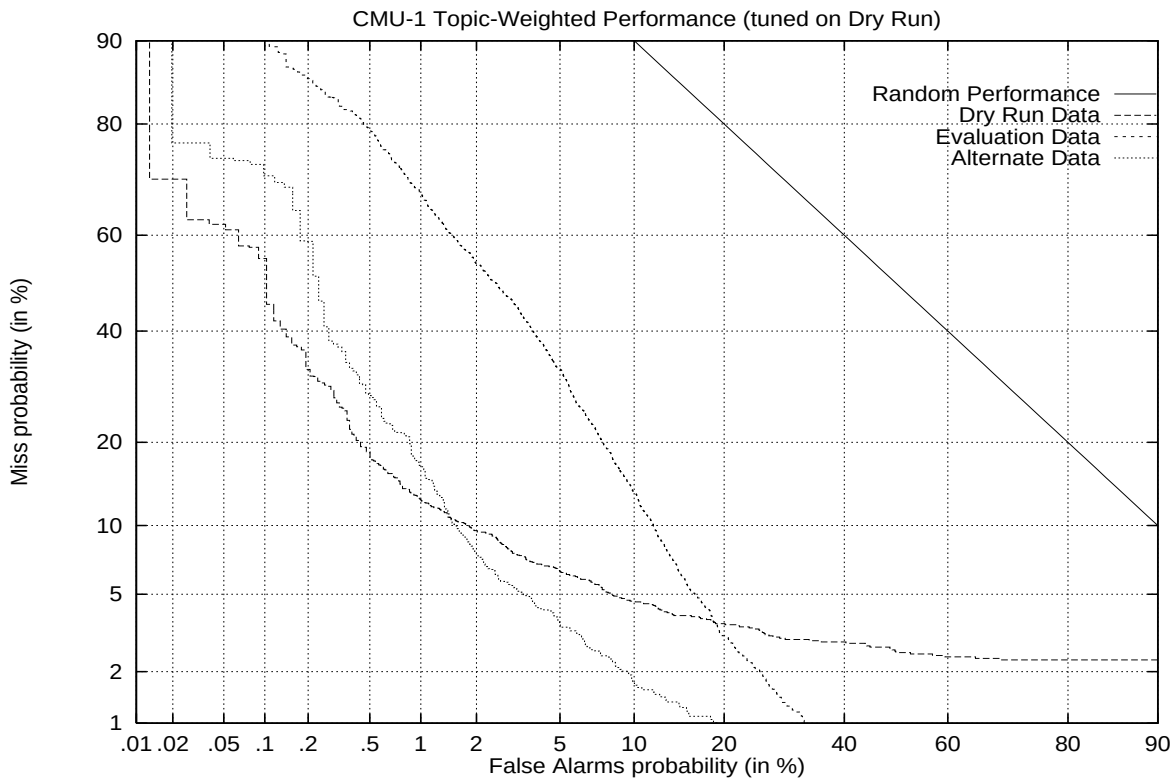

Figure 1: Performance Variation by Data Set: CMU-1 System

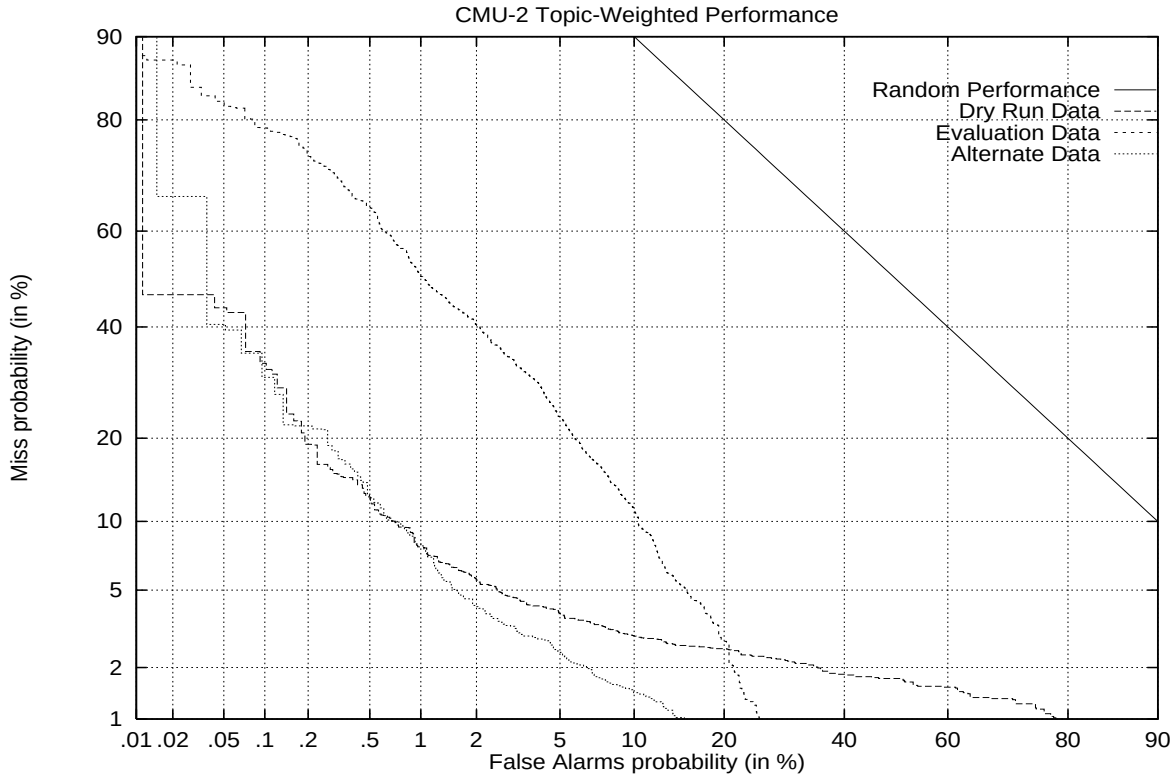

Figure 2: Performance Variation by Data Set: CMU-2 System

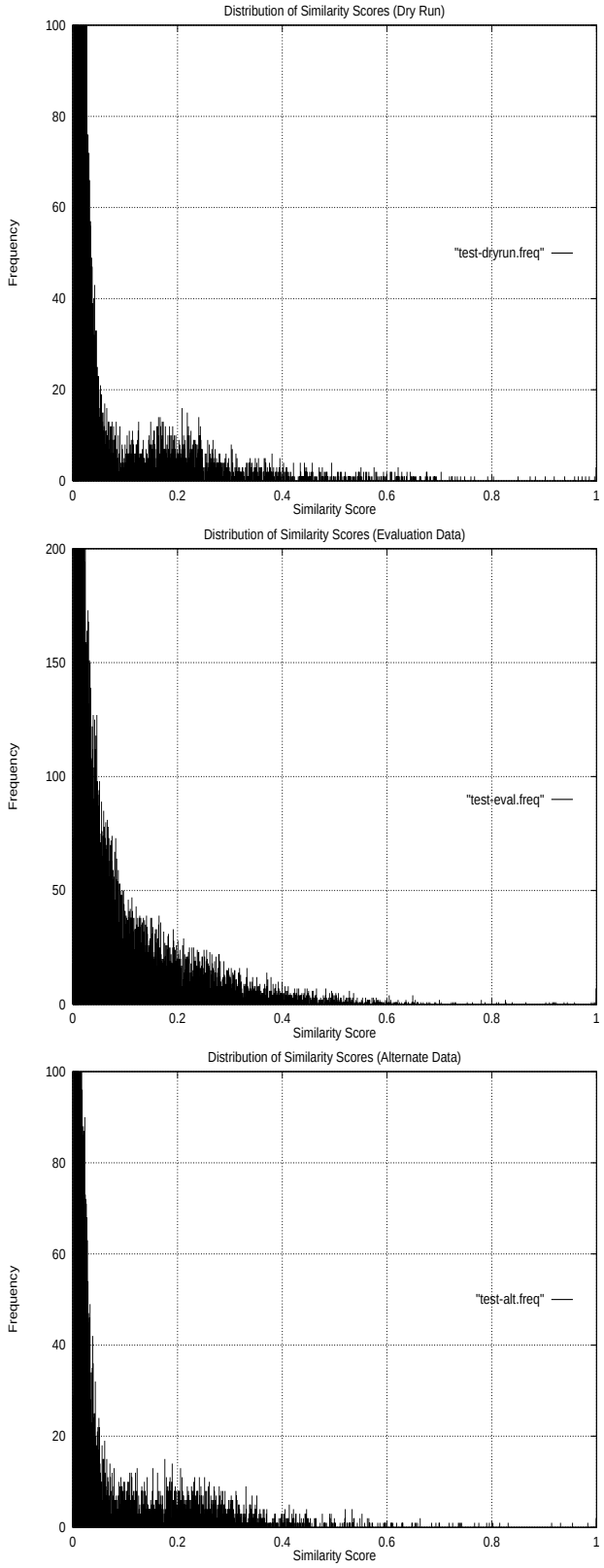

Figure 3: Comparing the Distributions of Similarity Scores

not clear whether there would be sufficient training material to accurately set the pairwise thresholds for more than three or four sources, even if the problem of dramatically different optimal thresholds between training and test sets were not an issue.

6. Acknowledgements

The authors would like to thank Tom Ault for his contributions to our Story Link Detection efforts.

References

1. Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D.Brown, Tom Pierce, and Xin Liu. CMU report on TDT-2: Segmentation, Detection and Tracking. In *Proceedings of the DARPA Broadcast News Workshop*, pages 117–120, San Francisco, CA, 1999. Morgan Kaufmann Publishers, Inc.
2. Jonathan G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of the 1997 IEEE ASRU Workshop*, pages 347–354, December 1997.
3. G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Pennsylvania, 1989.
4. Yiming Yang, Jaime Carbonell, Ralf D. Brown, Tom Pierce, Brian T. Archibald, and Xin Liu. Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems*, 14(4):32–43, July/August 1999. Special Issue on Applications of Intelligent Information Retrieval.