EXPLOITING TOPIC PRAGMATICS FOR NEW EVENT DETECTION IN TDT-2003

Ronald K. Braun and Ryan Kaneshiro

Stottler Henke Associates, Inc.
951 Mariner's Island Blvd., Suite #360
San Mateo, CA 94404

ABSTRACT

Stottler Henke participated for the first time in the New Event Detection (NED) track of TDT-2003 as a means of evaluating various prototyped components developed as part of a new story detection and topic tracking application. We combined a number of "pragmatics-based" classifiers in an ensemble-learning framework to identify the first story of a new topic and to link subsequent stories together as they unfold across multiple news streams. We present an overview of our techniques and a preliminary characterization of their performance based on our experimental runs for the TDT-2003 Evaluation.

1. INTRODUCTION

Stottler Henke is in the early stages of the development of a new story detection and topic tracking application called TOPIC ("Topic-Oriented Pragmatics and Invariant Chaining").* We postulate the existence of a variety of pragmatic processes and features that structure a news story as it unfolds over time. For each such feature that can be made computationally accessible, we implement a classifier that attempts the NED task using that feature as its basis for topic novelty judgment. These classifiers are housed in a committee architecture that applies an evidence combination technique to synthesize a global view of story novelty. Because an ensemble view of novelty is generated, no particular classifier need operate with perfect accuracy.

Though still in an early stage of development, we have implemented the framework architecture for the TOPIC system and have prototyped a number of classifiers and some simple evidence combination techniques. The TDT-2003 Evaluation (and, we hope, similar future evaluations) affords us an excellent real-world test-bed with which to characterize the performance of our system. Our focus up to this point has been written text news sources (AP newswire and New York Times articles) from the TDT-3 corpus. This evaluation is the first (rather eye-opening) attempt we have made to also process transcribed audio text (both automatically and manually generated) in conformance with the official evaluation test conditions [2]; as such, the techniques outlined herein may not be entirely applicable to transcribed audio and other non-text originating sources.

* This work is supported through DARPA SBIR contract DAAH01-03-C-R108.

The remainder of this section will give an overview of our approach. Section 2 describes the classifiers that appear in the TDT evaluation runs. Section 3 details some of the evidence combination techniques with which we have experimented. Section 4 concludes with a summary of our experimental results.

1.1. Pragmatics Framework

A fundamental premise underlying our work in pragmatics-based new event detection is that multiple structuring processes operate throughout the evolution of a story from the occurrence of events in the world to the reporting of those events to the consumption of resulting news stories by a target audience. These structuring processes may contain information that cues for story novelty and might thus be exploited by a NED system. We loosely define pragmatics as "non-semantic structure arising from *how* a topic is reported through time." With this focus we mean to avoid formal semantic modeling and a reliance on purely statistical linguistic techniques in an effort to bring to light other structuring aspects of news story text.

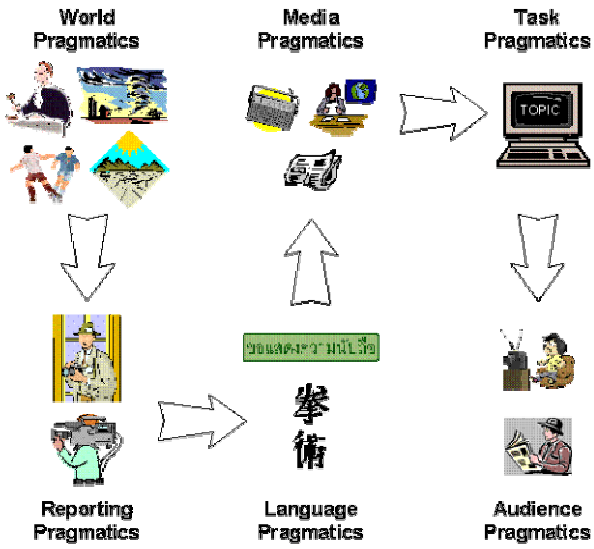

Figure 1. Pragmatics Framework guiding classifier development.

Figure 1 depicts the basic Pragmatics Framework that we use as an idea pump for generating classifiers. Space considerations

preclude a full discussion of each of the evolutionary stages of story development over time. Instead, we present a few potentially accessible features from each of the categories to give a flavor of the utility of this framework.

World Pragmatics. A story begins with the occurrence of some triggering event within the world. Events reported in a news story occur in the natural world and as such are subject to physical laws that structure them.

- Time / Location – Minimally, an event occurs at a particular time and place and lasts for some duration.
- Event actors – Events typically involve entities and may entail a transition between states associated with those entities. A particularly important class of entities is that of people, as events that make their way into the news are generally those that have effects on people.
- Causal effects – The TDT rules of interpretation identify highly correlated sequences of activities that co-occur for various topics of interest.

Reporting Pragmatics. A triggering event is observed by reporting agents who then summarize and contextualize the event within news reports.

- Newsworthiness – Since not all topics or all aspects of a given topic may be deemed newsworthy by a reporting agency, it may be possible to restrict the space of phenomena to be modeled (e.g., via topic category models) to a relatively small subset of all such phenomena. A reasonably broad range of coverage may thus be afforded by a highly circumscribed set of models.
- Necessary content – The aspects that make a story newsworthy may necessitate story content that can be correlated with topic novelty. For example, a phrase like "unknown number of victims" is likelier at the onset of a topic involving the deaths of people than later in the topic story stream.
- Linking aspects – When a new activity is reported for a topic, particularly when a large amount of time has passed since the story was last reported, reporting agencies will make explicit the linkage to a previous activity. It may be possible to differentiate these linking sentences from the general text of a new activity.
- Similarity of language use – Reporting agencies are likely to use the same references for entities across stories of a topic. An extreme case of this is the ossification of a phrase into a shorthand tag for the whole topic (e.g., "9-11"). The requirement for lexical novelty within news stories may be less than in other genres of writing.

Language Pragmatics. News reports are encoded within a particular language. Language is structured through grammatical and co-occurrence regularities.

- Word similarity – Related stories of a topic tend to share language features and may be recognized by such feature overlap. This is the basis for the success of full-text similarity topic tracking methods.
- Role of linguistic constituents – Verbs are used to express events and state changes within text at an abstract (categorical) level of generality; they are anchored to specific instances of an event by the nouns (e.g., entities) that serve as arguments to verb phrases.
- Analysis granularity – Phrase and sentence level analysis are appropriate for specifying individual, atomic events that compose the larger, thematic activities of a topic. The coarser-grained level of a story itself may tie together several disparate activities. Different levels of granularity may thus yield different features for analysis.

Media Pragmatics. News is conveyed in stereotypical formats defined by the media of presentation. Newswire stories may adhere to a pyramidal detail structure. Original written text sources have higher fidelity than transcribed or oral sources. Newspapers have space constraints that give rise to story dampening effects when events compete for print space.

- Source type features – Different sources may have different features available for exploitation. Location taglines, titles (as valuable summaries of content), audio teasers, and explicit linking newswire annotations may be extractable.
- Repetition of information – Information is repeated within a source in various ways. As a story evolves on a particular source, previous parts of the story may be reused or elaborated.
- Cross-source correlation – Information on one source may be predicted to appear on other sources. When this occurs, the intersection of information provided in both stories may contain the core elements needed to summarize and recognize the story.

Task Pragmatics. The TDT evaluation criteria provide a systematic set of biases that could be exploited to maximize performance with respect to the evaluation criteria

- Recognition of brief stories – In previous evaluation conditions, filtering out spurious links to brief stories would increase performance with respect to the evaluation metrics.
- Favoring high miss rates over high false-alarm rates – The error weights used in the evaluation may be optimized by tailoring the performance of classifiers to favor appropriate classes of error.
- Regularities in topic annotation – The topic annotation process itself might introduce regularities to the sorts of stories so annotated (e.g., the use of a search engine might ensure that one or more keyword phrases are present within all stories of that topic).

Audience Pragmatics. An operational topic tracking and novelty detection system operates as a filter of sorts between the reporting media and the target audience and must thus be tuned to the interests and intents of its audience. This pragmatics class is less relevant to the TDT evaluation task since audience interests are not likely to provide a priori structure to stories that can be exploited in gauging topic novelty.

The features from these pragmatic classes are computationally visible to varying degrees. Section 2 details some of the classifiers we have implemented thus far in an attempt to get at these features.

1.2. Evidence Combination

Each developed classifier exists to make a new-event detection decision with respect to each incoming story. Some means of combining the individual outputs of the collection is necessary to determine the final system NED verdict. Rather than training a single classifier over a range of features derived from the Pragmatics Framework, we have opted to house individual, one-feature classifiers in a committee-based architecture, utilizing committee-based or ensemble learning techniques to combine their results. New classifiers can be installed or removed at will and their contributions to the final system judgment evaluated.

Our evidence combination experiments are still preliminary in nature. We have evaluated classifier-independent techniques like majority voting schemes, which operate without knowledge of the individual constituents contained by the committee. We are also investigating classifier-aware methods, including Bayesian and regression techniques, which attempt to learn classifier weightings based on particular committee configurations.

2. NED CLASSIFIERS

In this section we will detail the seven classifiers that were included in our experimental runs for the NED task of TDT-2003. These are mostly variations on the full-text comparison methodology of traditional TDT systems, with each attempting to leverage a different pragmatic feature:

1. Vector Cosine (Baseline) – A full-text similarity technique in which each document is reduced to a Term Frequency / Inverse Document Frequency (TF/IDF) weighted feature vector of stemmed and stopped words. Vector cosine distance is used to gauge story similarity.

2. Temporal Weighted – A variation on the Baseline that uses a new weighting scheme we call Temporally-Weighted TF/IDF (TW-TF/IDF) weighting.

3. Linking Future – Future "events" are predicted to appear in upcoming stories; if a story satisfies a prediction, the story is deemed to be non-novel.

4. Linking Past – All previous stories are scanned for references to "events" contained in a new story. If an event match is found, the new story is deemed non-novel.

5. Naïve Tile – Stories are text-tiled and each story tile is compared to all previous stories using the Baseline methodology. Any tile which matches a previous story indicates a linking relationship.

6. Topic Conditioned – An incoming story is first placed into an activity cluster using the Baseline technique. The story's feature vector is then re-weighted with respect to the cluster and similarity is measured using this representation.

7. Activity Graph – Links together the clusters generated by the Topic Conditioned method and similarity is judged between the re-weighted stories in the connected clusters.

Ideally each classifier should use a judgment criterion that is orthogonal to those of other methods to increase the probability of a fruitful combination of evidence. It is not necessarily a requirement that a classifier be generally effective: if it covers some classes of data better than others or if it offers increased evidence for or against a certain judgment that leads to a better overall judgment, the classifier may still be an effective component of a committee. We therefore retain classifiers with overall poor performance if they offer some chance of combining well with other techniques (this is, of course, determined empirically).

2.1. The Baseline Technique

As our Baseline technique against which other classifiers are compared (and which can itself operate within a configured classifier committee), we have implemented a classifier, called Vector Cosine in our experimental runs, using the full-text similarity evaluation methodology developed in the early TDT evaluations (see, for example, [4]). Each story is tokenized, the tokens are stemmed and stopped, and a bag-of-words feature vector is constructed to represent the story. Each element of the vector is weighted according to its TF/IDF value (computed from the full TDT3 corpus). To evaluate the novelty of a new story, the distance of its vector from previous story vectors is determined using vector cosine difference; if the cosine distance is less than a threshold, the new story is deemed to match the previous story and is judged to be non-novel.

2.2. TW-TF/IDF Weighting

Given the pragmatic observation that all triggering events for a news story occur at particular place and time, we hypothesized that most descriptions of key events (from a topic-tracking point of view) within a news story would have a temporal component. Temporal cues should thus be rather predictive of important swaths of text within a story. Without resorting to actual temporal modeling and reference resolution (see [6] for a more sophisticated treatment of temporality), we developed three classifiers that attempted to leverage temporal cues to useful effect. Each of these uses the notion of a *temporal reference* to target parts of text for processing. A temporal reference is any word that identifies a particular point in time. Figure 2 lists the fifty-one temporal references currently in use within our system.

tomorrow, day, month, week, year, January, Jan, Febuary, February, Feb, March, April, Apr, May, June, Jun, July, Jul, August, Aug, September, Sep, Sept, October, Oct, November, Nov, December, Dec, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, yesterday, morning, noon, evening, night, summer, winter, autumn, spring, fall, o'clock, dawn, dusk, midnight, today

Figure 2. Temporal references.

Since the text around temporal references is hypothesized to be of greater value in gauging the similarity of two stories, we devised a modification to traditional TF/IDF weighting called TW-TF/IDF. The essential idea is to increase the weighting for words depending on their proximity to a temporal reference. To do this, we identify all of the temporal references within a story and assign each word a distance value (*d*, in the equation below) equal to the number of stopped words away from its closest temporal reference. A Gaussian (Bell) distribution is situated atop each temporal reference yielding a temporal weighting contribution in the interval [0.0, 0.4]. The width of the curve (that is, the number of words that constitute one standard deviation) is a parameter of the classifier and is represented by *s* (the scale factor) below; we find that a scale factor of s = 15 stopped words is reasonably optimal. All features (i.e., word stems) receive a temporally-adjusted TF weight (*twtf*) determined by summing the contributions of each instance (*i*) of the feature in the story as defined in the following formula:

$$twtf\,(feature) = \sum_{\forall i}\left(0.4 + \frac{e^{-\frac{d^2}{2s^2}}}{\sqrt{2\pi}}\right)$$

While TF weighting typically assigns a base weight of 1.0 per instance, we use of base of 0.4 to match the magnitude of the Gaussian contribution. Thus, temporal words (assuming they aren't stopped) would have a *twtf* weight contribution of 0.8 per instance, while all words in a story containing no temporal references would have a *twtf* weight contribution of 0.4 per instance. The intuition is that words around temporal references are in some sense approximately (as decayed by the Gaussian) twice as useful as non-temporally associated words. The IDF weight component is computed as per normal TF/IDF weighting.

If the assumption that temporal references are good content cues is correct, we would predict that inverting the weighting for temporal references (we'll call this Anti-TW-TF/IDF weighting) should degrade performance. To invert the weighting, we subtract the Gaussian contribution from 0.4, such that temporal references are weighted 0.4 (0.4 base + 0.0 Gaussian) per occurrence for the *twtf* component and terms distant from such references are weighted 0.8 (0.4 base + 0.4 Gaussian) per occurrence.

Figure 3 shows the performance of the Temporal Weighting classifier using both TW-TF/IDF and Anti-TW-TF/IDF weighting on approximately 28000 stories from the TDT3 corpus. (This subset consists of all AP and NYT stories plus another 10,000 or so from the manually transcribed audio sources chosen to reflect the same story per source density found in the TDT4 corpus. In the remainder of this paper, this dataset will be referred to as the MiniManual dataset.) While the difference in performance is relatively small, the Anti-TW curve is fairly consistently above the TW curve. A plot of the Baseline is sandwiched by the two curves. This effect is generally consistent across data sets.

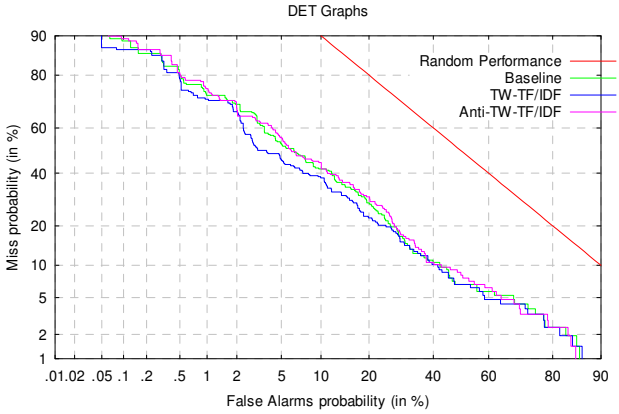

Figure 3. DET curves for TW-TF/IDF and Anti-TW-TF/IDF.

Table 1 shows the optimal topic-weighted CFSD scores (see [2] for a definition and discussion of DET curves and CFSD scoring) for each of the three versions of the classifier over the MiniManual dataset. The gain over the Baseline is approximately 10%, though our experience with a variety of other data partitions would locate this value closer to 4% on average; the fact of an increase does appear to be a consistent phenomenon. (The MiniManual dataset is slightly anomalous in that it is generally the case that the anti-TW CFSD value is greater than the baseline value, which is in turn greater than the TW value.) Whether the slight accuracy increase is worth the computational cost of the TW-TF/IDF weighting is certainly a debatable issue. Of greater interest is the disparity between the Anti-TW version and the TW version, which supports the hypothesis that temporal references are good cues to exploitable content-structuring processes.

TW-TF/IDF	TF/IDF (Baseline)	Anti-TW-TF/IDF
0.6705	0.7537	0.7414

Table 1. Optimal topic-weighted CFSD scores.

2.3. Event Linking

If temporal references do cue for text describing particular physical events in a news story, it is feasible that the primary entities and actions that characterize the event (e.g., fill its argument slots in a frame-based representation of the event) will be found in close proximity to the temporal reference. These physical events may comprise the primary events of the topic being reported; recognizing references to the same event across

stories should thus aid in identifying related stories with respect to an evolving topic. A potentially important class of this phenomenon is to be found in what we call *linking events*. A linking event is a brief description of an event in a new story that explicitly evokes the larger topic under consideration. When a topic has been dormant for some period of time, new stories that emerge on the topic almost always mention a linking event to make the relationship of the story to previous stories explicit for the audience. For example, a story about the Pope encouraging Catholics to donate to a relief fund after a hurricane may have only a single sentence about the actual hurricane (this will almost certainly contain a temporal reference), while the rest of the text is quite novel in terms of word similarity to previous hurricane stories; this sentence makes explicit the event that contextualizes the story. Recognizing such an event can thus in theory be quite important in assigning the story to an existing hurricane topic.

To try and get at the granularity of physical events in a story, we make use of temporal references as cues to such an event. For the purposes of the next two classifiers to be discussed, we define an *event reference* in text to be any sentence that contains a temporal reference, plus some number of context sentences on either side of that sentence (we find two sentences on each side to be optimal).

The event references of a story are thus the set of sentence clusters situated around temporal references in the story. We recast the novelty detection problem as a reference resolution problem by treating all event references as potentially referring to previous or future stories. Under this formulation, we developed two classifiers called Linking Past and Linking Future.

The Linking Past classifier operates as follows. For each story to be judged for novelty, all of the event references for the story are determined. A feature vector is constructed for each event reference (using the Baseline TF-IDF methodology) and this vector is cosine-compared to all previous whole stories encountered. If a threshold is met, the reference is assumed to refer to the topic of the story and the reference is resolved. If a story has any resolved event references, it is considered to be not novel; stories without event references or with no resolved references are considered to be first stories in a new topic.

The Linking Future classifier operates in a similar manner, except each event reference in a story is considered to be a prediction that the event will occur in the future and will thus be described by an upcoming story. For each story processed by the system, all of its event reference feature vectors are added to a prediction list maintained by the classifier. Each new story is cosine-compared against this list of predicted events. If the new story matches any predicted event, the story is judged not novel; a story that satisfied no predictions is considered to be a first story.

TF/IDF (Baseline)	Linking Past	Linking Future	Past + Future Majority
0.5743	0.6192	0.5766	0.5477

Table 2. Optimal topic-weighted CFSD scores for event linking classifiers.

The Linking Past and Linking Future classifiers typically complement each other rather well. Table 2 details the optimal CFSD values for the two linking event classifiers, the baseline, and a majority committee that combines both classifiers. The data set used was a subset of MiniManual consisting of the full set of AP and NYT newswire stories only. Unfortunately, these two classifiers can not handle audio transcribed materials (and thus perform poorly on the full MiniManual dataset), as some property of audio transcriptions confounds their techniques. If we can identify that property and filter it out, we may be able to generalize the classifiers. Figure 4 shows the DET curves associated with the classifiers on the same data set.

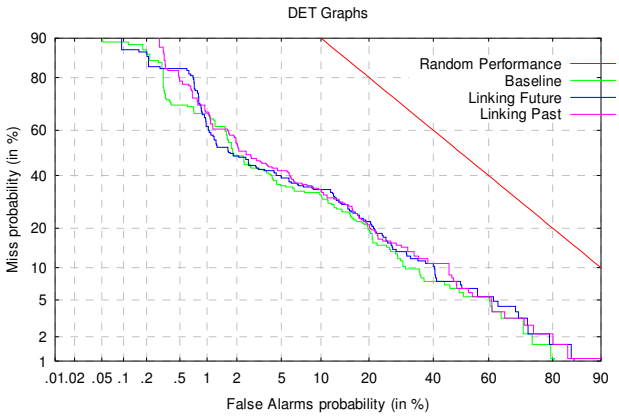

Figure 4. DET curves for event linking classifiers.

2.4. Topic Granularity

Many stories span multiple topics or are comprised of several sections that are each rather disparate from the other. Recognizing that full news stories may be too coarse-grained when trying to match topical text in a story to same-topic text in another story, we implemented a classifier called Naïve Tiling that applies Marti Hearst's text tiling algorithm [1] to a story, thereby fragmenting it into multiple tiles. The classifier is considered naïve in the sense that we make the simplifying assumption that each tile in the story potentially pertains to a separate topic. The Baseline text similarity methodology is then applied at the tile level, rather than at the story level. For a new story to be classified, each of the tiles in the story is compared to all previous stories; if a match is found, the new story is considered to be not novel, otherwise it is judged a first story.

The smaller granularity of a tile produces more spurious links between stories because there are fewer words which need to overlap in order to produce a high similarity score. As a result the miss rate tends to increase along with the number of previously seen stories. A temporal locality assumption is made to offset this. A five week sliding window is used when considering a new story (this window size was determined empirically). A new story is compared against only those stories within the window. This technique decreases the miss rate but increases the possibility of a false alarm with a story that links back to one that is outside of the window.

Figure 5 shows the performance of the Naïve tile classifier along with the Baseline on the MiniManual dataset.

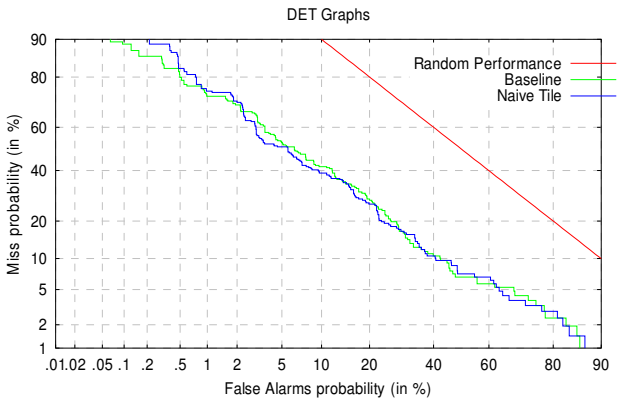

Figure 5. DET curve for the Naïve Tile classifier.

Table 3 shows the optimal topic-weighted CFSD for the Naïve Tile classifier and the Baseline over the same story set. The optimal CFSD values for the Naïve tile algorithm show a modest 5% improvement over the Baseline. The values are slightly better across other story sets although the DET curves typically intertwine.

Naïve Tile	TF/IDF (Baseline)
0.7162	0.7537

Table 3. Optimal topic-weighted CFSD scores for the Naïve Tile classifier.

2.5. Activity Clustering

Stories that discuss similar categories of events can be erroneously linked together by the Baseline technique. Within a particular category, word usage remains similar regardless of the specific event. These overlapping words are informative when compared against the global spectrum of stories but are not so useful when examining stories within the same activity. For example, stories about hurricanes tend to contain words like "hurricane", "wind", "debris", etc. Stories about Hurricane Mitch can be incorrectly considered similar to stories about Hurricane Georges because of the words common to the hurricane topic.

The Topic Conditioned classifier is a variation on the first story detection system developed at CMU [8]. TF/IDF weights for a story are initially calculated using all of the stories in a training set as the document collection. The story is then placed into an activity cluster and the weights are recalculated using the cluster members as the document collection. Instead of constructing topic clusters (clusters of all of the stories in a topic) as CMU does, we construct activity clusters (clusters of all of the stories in an activity of a topic). After re-weighting the story vector, it is compared against the other cluster members using the cosine similarity metric. If a match is found within the cluster, the new story is not considered novel, otherwise it is a new event. The re-weighting is designed to decrease the weight of topic-specific words ("hurricane", "wind", "debris") and increase the weight of the instance-specific words ("mitch", "georges").

The CMU system used a manually generated set of topics and trained a classifier to place stories into the appropriate cluster. We are attempting to use automatically generated topic activities rather than picking them in advance. Each incoming story is compared against the centroid of existing activity clusters. If the similarity between the story and a centroid exceeds a threshold, the story is added into the cluster, otherwise it becomes the initial member of a new cluster and will be considered a new event.

Proper nouns drown out the other words in a story because of their higher TF/IDF weights. The resulting topic clusters tend to be centered around people and locations rather than activities. To work around this, each word in a story is assigned a part of speech tag. Words that are marked as a proper noun are removed from the story before comparing against the clusters. The proper nouns are added back into the story vectors before the re-weighting occurs.

The Activity Graph classifier attempts to leverage the fact that a single topic comprises stories from disparate activities. A natural disaster event may consist of stories about the initial disaster, the recovery efforts, and the rebuilding process. Stories describing the different activities may have little word overlap. With the Topic Conditioned model, there is no way to link two stories together once they have been placed into different clusters. The Activity Graph classifier treats each of the clusters as nodes in a fully connected graph. The intra-cluster similarity is computed in the same fashion as the Topic Conditioned model. An inter-cluster similarity value is generated by comparing the re-weighted story against the re-weighted stories in each of the other clusters. The final confidence value is a combination of the inter-cluster and intra-cluster similarity values. Because each cluster increases the weight of the instance-specific words, stories which were initially dissimilar can be linked using the boosted weight of the words that characterize a particular event instance.

Topic Conditioned	Activity Graph	TF/IDF (Baseline)
0.9209	0.8207	0.7537

Table 4. Optimal topic-weighted CFSD scores for activity clustering classifiers.

Table 4 shows the optimal topic-weighted CFSD for the Topic Conditioned, Activity Graph, and Baseline classifier over the same story set. The Activity Graph model does better than the Topic Conditioned classifier mainly because it compares the incoming story against all of the previous ones rather than just the stories within a single cluster. If the Topic Conditioned model chooses the wrong cluster, there is no way to recover. For both the Topic Conditioned and Activity Graph classifiers, each incorrectly classified story pollutes the cluster centroid and makes it easier for the focus of the cluster to drift. Despite the problems associated with choosing the correct cluster, the Activity Graph classifier contributes value to new event detector committees.

Figure 6 shows the performance of the Topic Conditioned. Activity Graph, and Baseline classifier on the MiniManual dataset.

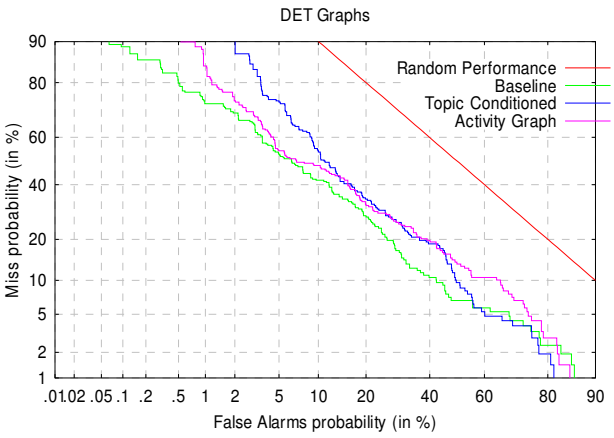

Figure 6. DET curves for the activity clustering classifiers.

3. EVIDENCE COMBINATION

Once a committee of independent classifiers (each making a separate NED story novelty judgment) has been assembled, their results must be combined in a systematic manner. We've investigated some preliminary techniques that can be broken into two general categories: classifier-independent and classifier-aware. Classifier-independent strategies are generic strategies that operate using a criterion independent of the contents of the committee. Our initial foray into this genre has focused on majority and authority voting schemes. Classifier-aware strategies attempt to learn an optimal combination method based on the specifics of the committee configuration. For this latter class, we have experimented with a naïve Bayesian approach and a linear combination methodology. These latter methods generally attempt to learn weights or weight classes for the contribution of each classifier to a committee's final judgment.

3.1. Majority Voting

With majority voting schemes, the members of the committee are each polled for a NED judgment and the majority decision is taken as the system decision. The intuition is that the more independent perspectives are in agreement on a judgment for a story, the more likely that judgment is to be correct.

Each classifier can be trained to minimize its topic-weighted CFSD score on a training set or to globally minimize all classes of errors in general. Our experiments with evidence combination are quite preliminary and we have as yet to determine a systematically best practice for configuring individual classifiers. For the TDT-2003 evaluation, we trained classifiers to minimize the following quantity: total *miss errors* + total *FA errors* – total *brief errors* + total *lucky errors*. Miss errors occur when the classifier says the story being evaluated is not novel when in fact it is; FA errors occur when the story being evaluated is judged novel when it in fact is not; brief errors occur when a story links to a brief story (a story wherein less than 10% of content is on-topic) and so is deemed not novel when it should be judged novel (this is according to previous TDT evaluation criteria); and lucky errors occur when a story is correctly labeled not novel because it links to a previous *off-topic* story (i.e., the classifier made a correct judgment but did so using erroneous reasoning). The brief error rates are subtracted out of the quantity instead of added to compensate for the fact that such stories are penalized under previous TDT evaluation criteria but not in the current test condition specification.

The confidence generated by the system is the average normalized distance between each classifier's independent confidence value and its decision threshold. In the case of ties (a frequent phenomenon in even-numbered committees), the maximal average normalized difference between the novel versus the non-novel voters decides the system.

3.2. Authority Voting

In an authority voting committee, a single classifier is specified as the primary classifier and its judgment is the default decision. The primary classifier is typically the one that is deemed the best overall performer. (Currently, we use the Temporal Weighted classifier for this.) Other classifiers are allowed to override the default judgment if their expertise allows them to say with near certainty that a story is not novel. The intuition here is that we use the best performer of the bunch unless another classifier is extremely sure of its answer, in which case it is allowed to override the default decision and correct some of its presumed errors. We only allow not-novel overrides because of the asymmetry inherent in topic novelty determination: it is generally possible to make an instant determination of non-novelty given a new story based on some criteria, but not possible to make an instant novelty determination (the absence of evidence is not generally useful as evidence of topical absence).

All other members of the committee are trained to minimize false alarm errors (since these are the ones that dominate the CFSD determination). Again, it is not clear that this is the best practice. For the TDT-2003 evaluations, we trained each committee member other than the primary classifier down to a 2% false alarm error rate. When the committee is handed a story for evaluation, all authorities evaluate the story for non-topicality. If any of them decide not-novel, that judgment is returned, otherwise the default decision is returned. The reported confidence for a non-novel judgment is the maximal relative difference from each individual classifier's judgment threshold over all classifiers that voted non-novel; for a novel judgment, we used the minimal relative difference from each threshold over all classifiers that gave a novel judgment.

3.3. Naïve Bayes

The Naïve Bayes algorithm has been shown to be an effective meta-classifier for stacking individual committee members together [5]. With this technique, the hard yes/no decisions output by each of the individual classifiers in the committee are

used as feature values. We assume independence between classifiers and the best individual thresholds are used for each member. The meta-classifier outputs a class distribution containing the likelihood of the story being a new event or not. The class with the greater likelihood is used as the hard decision and the final confidence value is defined as:

$$\frac{P(yes)}{P(yes)+P(no)}$$

There are a couple of drawbacks to using the hard decision of the committee members. One is that the combiner is completely dependent upon the threshold choice. A bad choice of threshold will yield poor results for the combiner. Another problem is the discreteness of the confidence values that are calculated. The Naïve Bayes class distribution is essentially a combination of the hard decisions made by the individual committee members. There are only two possible values for each committee member (yes and no) so a committee with 3 members can have only 8 possible confidence values. This skews the DET curve because once a threshold is set to be less than one of these discrete values, every single story with that confidence value becomes a first story. This causes the false alarm rate to suddenly jump when the miss rate is lowered.

3.4. Linear Combination

The "best overall results generator" (BORG) scheme used at CMU [7] takes a linear combination of the normalized confidence values for each of the committee members. The confidence values for each member in the committee are normalized using the formula:

$$x' = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation calculated over a training set. The normalized values are summed together and the resulting value is normalized once more to produce the final committee judgment.

4. EVALUATION RESULTS

We submitted five runs each for the two evaluation conditions (manual and ASR) of the TDT-2003 NED track. (The final adjudicated results can be found at [3].) Due to the preliminary nature of our evidence combination methodology and the desire not to flood NIST with an obnoxious number of runs to score, we somewhat arbitrarily picked a few different committee configurations that we felt had promise. All classifiers were trained on the MiniManual dataset described in Section 2.2; given a lack of familiarity with the TDT4 corpus used for evaluation, we hoped this set might conform reasonably to the characteristics of that target corpus.

NIST Code	Combo Method	Classifiers	Optimal CFSD
SHIA1	Majority	Naïve Tile Linking Future Temporal Weighted	0.6346
SHIA2	Authority	Naïve Tile Linking Future Temporal Weighted	0.6633
SHIA3	None	Temporal Weighted	0.6383
SHIA5	Naïve Bayes	Activity Graph Linking Past Linking Future Vector Cosine	0.7197
SHIA6	Linear Combo	Vector Cosine Linking Future	0.6399

Table 5. ASR condition evaluation runs.

Table 5 shows the committee composition, evidence combination strategy, and final optimized CFSD number for each of the evaluation runs we submitted for the official ASR transcription evaluation condition. Table 6 similarly covers the manual transcription condition.

NIST Code	Combo Method	Classifiers	Optimal CFSD
SHIA2	Authority	Naïve Tile Linking Future Temporal Weighted	0.6554
SHIA3	None	Temporal Weighted	0.6574
SHIA4	Majority	Activity Graph Linking Future Naïve Tile Temporal Weighted Topic Conditioned	0.6463
SHIA5	Naïve Bayes	Activity Graph Linking Past Linking Future Vector Cosine	0.6578
SHIA6	Linear Combo	Vector Cosine Linking Future	0.6797

Table 6. Manual condition evaluation runs.

The SHIA3 condition can be viewed as a baseline condition of sorts insofar as it represents the best single classifier (i.e., Temporal Weighted) out of all of those developed. We would hope that committees would generally outperform this baseline. On the surface, committee results are somewhat disappointing with respect to the optimal CFSD value. However, there are several observations that give us some hope that we can push these techniques significantly further.

The first is illustrated by an inspection of the DET curves for the manual condition shown in Figure 7. (Unfortunately a clean ASR version with only our runs is not available and the all-sites composite for the ASR condition is not easily readable.) While the optimal CFSD of the five classifier committee SHIA4 is only marginally more optimal than the baseline (0.6463 vs. 0.6574), the DET curve has the nice characteristic that it fairly consistently runs beneath the SHIA3 baseline by an appreciable margin. This is a nice result insofar as it was quite reasonable to suppose that a large number of classifiers trained and tuned to the TDT3 corpus

would potentially amplify each other's errors rather than compensating for them. Instead, multiple viewpoints do seem to reduce the variance and error over the constituent classifiers and yield a more coherent result. This offers some preliminary support for the hypothesis that evidence combination is a useful tool when performing the NED task.

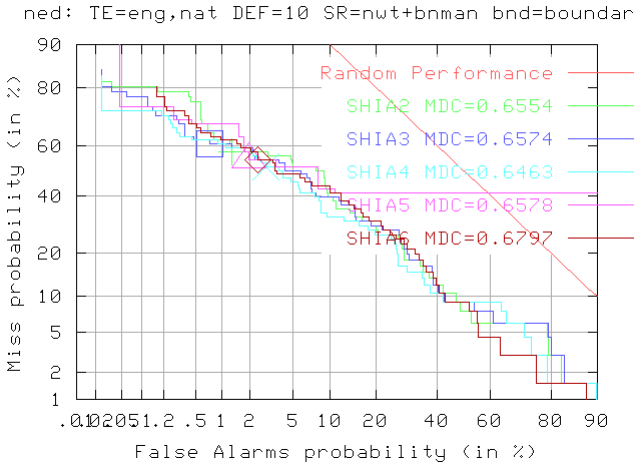

Figure 7. DET curves for the manual condition.

Another mitigating factor is that all of these classifiers were developed based on our work with written sources only (i.e., the AP newswire and the New York Times); this evaluation was our first experience with transcribed audio sources, which have rather significantly different characteristics than written sources. We know that the Linking Past and Linking Future classifiers operate quite poorly over a mix of audio and written sources, but operate reasonably well over written-only sources. Assuming a methodology can be found to allow varying classifiers to judge only for specific sources in the final mix, we should be able to considerably improve the quality of the data being combined. We would also expect that a different set of classifiers might be developed to address audio source characteristics directly.

One assumption of our work is that the more orthogonal the criterion for judging a story as novel or not used by different classifiers, the more fruitful one might expect the combined view of the problem to be, insofar as the classifiers aren't essentially "doing the same thing." Most of the classifiers we have developed thus far are predicated on the methodology of feature vector distance for their primary judging criteria, though each classifier typically applies some pragmatic slant to the problem. As we develop additional pragmatic classifiers, we hope to move in more divergent directions that should afford greater orthogonality to their judgment criteria, hopefully permitting a more effective fusion of evidence within a given committee.

The TDT-2003 evaluation has proven to be an illuminating experience and an excellent test-bed in which to validate some of the ideas underlying the development of an operational first-story detection and topic-tracking application. We look forward to participation in TDT-2004 where we hope to have a more mature version of the techniques described herein available for evaluation.

REFERENCES

1. Hearst, M. "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages", *Computational Linguistics*, 23 (1), pp. 33-64, March 1997.
2. National Institute of Standards and Technology (NIST). "The 2003 Topic Detection and Tracking (TDT2003) Task Definition and Evaluation Plan", http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm, 2003.
3. National Institute of Standards and Technology (NIST). "Topic Detection and Tracking 2003 Evaluation", http://www.nist.gov/speech/tests/tdt/tdt2003/tdt2003_official_results_20031029, 2003.
4. Schultz, J. M. and Liberman, M. "Topic Detection and Tracking using idf-Weighted Cosine Coefficient", *Proceedings of the DARPA Broadcast News Workshop*, pp. 189-192, 1999.
5. Seewald, A. K. "Exploring the Parameter State Space of Stacking", 2002 IEEE International Conference on Data Mining, pp. 685-688, 2002.
6. Swan, R. and Allan, J. "Automatic generation of overview timelines", Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 49-56, 2000.
7. Yang, Y., Ault, T., and Pierce, T. "Combining multiple learning strategies for effective cross validation", Proceedings of ICML-00, 17th International Conference on Machine Learning, pp. 1167-1182, 2000.
8. Yang, Y., Zhang, J., Carbonell, J., and Jin, C. "Topic-conditioned Novelty Detection", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 688-693, 2002.