The University of Michigan at TREC 2003

Jahna Otterbacher, Hong Qi, Ali Hakim and Dragomir Radev
University of Michigan
Ann Arbor, MI 48109
{jahna, hqi, alihakim, radev}@umich.edu

February 2, 2004

1 Introduction

This year the University of Michigan team participated in all four tasks of the Novelty track. Our basic approach for all the tasks involved using our multi-document summarizer, MEAD [1], as well as Maxent 2.1.0 software for training maximum entropy classifiers[1]. We submitted five runs for each of the four novelty tasks.

2 General Approach

2.1 Data

We created training and dev/test data sets for training models for each of the four tasks. In tasks 1 and 2, we used the Novelty 2002 test data for training. 25 clusters were assigned to our training set, while the remaining 25 were used for our dev/test data. For tasks 3 and 4, we were given some information about the 2003 test data, so our training and dev/test data were based on the information provided. Table 1 describes our training, dev/test and test data for each of the four tasks.

2.2 Features used in classification

We experimented with six sentence features in building classifiers to detect novel and relevant sentences. Three were standard sentence features in the MEAD package while the other three used the topic queries provided for each cluster in the data.

- **Centroid** The centroid score quantifies the extent to which the sentence contains lexical items that are key to the overall cluster of documents.

[1] *http://maxent.sf.net*

Task	Description	Training data	Dev/Test	Test
1	Find all novel and relevant sentences	25 clusters from 2002 data	25 clusters from 2002 data	50 clusters from 2003 data
2	Given the list of relevant sentences for all documents in the test data, find all novel sentences	Relevant sentences from 25 clusters in the 2002 data	Relevant sentences from 25 clusters in the 2002 data	50 clusters from 2003 data
3	Given the list of relevant and novel sentences in the first 5 documents of all 2003 clusters, find the relevant/novel sentences for the last 20 documents	Relevant and novel sentences for the first 5 documents in clusters N1-N25	Relevant and novel sentences for the first 5 documents in clusters N26-N50	Last 20 documents of the 50 clusters in the 2003 test data
4	Given the list of relevant sentences in all documents in the test data and the list of novel sentences in the first five documents, find the novel sentences in the last 20 documents	First 5 documents in clusters N1-N25	First 5 documents in clusters N26-N50	Relevant sentences in the last 20 documents of the 50 test clusters

Figure 1: Data used in each of the four tasks

- **Length** The length of the sentence by number of words.
- **Position** The position of the sentence in its original document.
- **QueryTitleCosine** The cosine o the vectors representing the title portion of the query for the cluster and the sentence.
- **QueryTitleWordOverlap** The word overlap between the title portion of the query for the cluster and the sentence.
- **QueryNarrativeWordOverlap** The word overlap between the narrative portion of the query of the cluster and the sentence.

For each task, the features were discretized. The cosine feature was broken down into three categories (high, medium and low scores), while the remaining five features were converted to be binary. The threshold values for discretizing the feature scores were tuned using the training data for each task.

2.3 Model

As mentioned, we built maximum entropy classifiers to find relevant and novel sentences in the four tasks. In other words, for each task we found the best conditional exponential model to determine the probability of a sentence being relevant (or novel), given the values of its six features.

- The probability of a sentence being labeled (as novel or relevant, depending on the task) is

$$Prob(label|f) = Z * e^{\sum_i \lambda_i * fd_i(lf)}$$

where w_i are the model parameters (weights), Z(f) is a normalizing factor independent of l, and $\lambda_i = log(w_i)$.

For each task, we trained the above model in order to find the probability of all of the sentences in the data being assigned a label of relevant or novel. We then ranked the sentences by their respective probabilities. Finally, we experimented with our threshold values (the percentage of top ranked sentences to submit for the run). To summarize, the approach we used this year involves four steps for each task:

Step 1 Using the appropriate training data for the task, tune the threshold values in our discretization process for the sentence features.

Step 2 With the discretized data set, train the maximum entropy classifier for finding relevant/new sentences as appropriate to the task.

Step 3 Use the model to predict the probabilities of the sentences in the test data being assigned relevant/new labels. Rank the sentences by their probabilities.

Step 4 Determine the optimal cut-off value for submitting sentences in their ranked order, to the list of relevant or new sentences.

3 Task 1

In the first task, our goal was to identify the relevant and novel sentences in the 50 test clusters, given no additional information. Therefore, we used 25 clusters from the Novelty 2002 test data for our training and the other 25 for development. We developed one discretization procedure for this task, and focused on tuning our threshold for the percentage of top-ranked sentences to submit.

Table 2 shows the five runs we submitted, the performance (F-measure) we obtained on our dev/test data set, as well as the official F-measure from NIST. Note that the F-measure is reported as the average score over all clusters evaluated.

4 Task 2

For task 2, in which the list of relevant sentences was given for all 25 documents in each cluster, the goal was to identify all of the novel sentences in all documents. Therefore, a naive baseline is to submit all of the relevant sentences as novel. This naive approach achieved an F-measure of 0.6174, which seemed rather high. Therefore, we experimented both with our feature discretization procedure as well as with tuning our cut-off threshold. Our submitted runs include different combinations of discretizations and cut-off thresholds. They are described in Table 3.

Run	Description	F-measure on dev/test data	Official F from NIST
Umich11	top 3% of ranked sentences	0.113 (rel) 0.117 (new)	0.192 (rel) 0.182 (new)
Umich12	top 3.5% of ranked sentences	0.126 (rel) 0.119 (new)	0.086 (rel) 0.093 (new)
Umich13	top 4% of ranked sentences	0.129 (rel) 0.121 (new)	0.110 (rel) 0.115 (new)
Umich14	top 4.5% of ranked sentences	0.127 (rel) 0.119 (new)	0.134 (rel) 0.136 (new)
Umich15	top 5% of ranked sentences	0.125 (rel) 0.118 (new)	0.156 (rel) 0.155 (new)

Figure 2: Task 1 - Runs submitted, F-measure on dev/test data, and official F-measure from NIST

Run	Description	F-measure on dev/test data	Official F from NIST
Umich21	Top 95 % of ranked sentences Discretization 2	0.617	0.394
Umich22	Top 90% of ranked sentences Discretization 2	0.592	0.377
Umich23	Top 85% of ranked sentences Discretization 2	0.545	0.361
Umich24	Top 95% of ranked sentences Discretization 1	0.599	0.209
Umich25	Top 90% of ranked sentences Discretization 1	0.596	0.212

Figure 3: Task 2 - Runs submitted, F-measure on dev/test data, and official F-measure from NIST

5 Task 3

In task 3, we were given the lists of relevant and novel sentences from the first 5 documents of each of the 50 clusters in the 2003 test data. Using no additional information (such as the list of all relevant sentences we were given in task 2), the goal was to identify all relevant and novel sentences in the remaining 20 documents of each of the 50 clusters. We decided to use the first 5 documents in clusters N1-N25 of the 2003 data as our training data set, and the first 5 documents in clusters N26-N50 as our dev/test set for this task. In retrospect, we should have also included data from the 2002 competition in our training and dev/test sets. This is because we didn't know if there were any significant differences between the first 5 and last 20 documents in the 2003 clusters that might have influenced the prior probabilities of a sentence taking one of the two (relevant/new) labels.

The prior probability of being a relevant or novel sentence was quite high in our training data set. Therefore, we considered the naive baseline of submitting all sentences as being relevant and novel. On our dev/test set, this naive approach achieved an F-measure of 0.54 for novel sentences and 0.67 for relevant sentences. By using one discretization procedure and focusing on tuning our cut-off threshold on the ranked sentences list, we found that we could beat these baselines. Table 4 shows our results.

Run	Description	F-measure on dev/test data	Official F from NIST
Umich31	top 80% of ranked sentences top 65% of ranked sentences	0.72 (rel) 0.61 (new)	0.560 (rel) 0.409 (new)
Umich32	top 75% of ranked sentences top 75% of ranked sentences	0.71 (rel) 0.59 (new)	0.565 (rel) 0.405 (new)
Umich33	top 70% of ranked sentences top 80% of ranked sentences	0.71 (rel) 0.59 (new)	0.569 (rel) 0.399 (new)
Umich34	top 65% of ranked sentences top 90% of ranked sentences	0.71 (rel) 0.56 (new)	0.566 (rel) 0.385 (new)
Umich35	top 90% of ranked sentences top 60% of ranked sentences	0.68 (rel) 0.58 (new)	0.543 (rel) 0.408 (new)

Figure 4: Task 3 - Runs submitted, F-measure on dev/test data, and official F-measure from NIST

Run	Description	F-measure on dev/test data	Official F from NIST
Umich41	all relevant sentences in last 20 documents	0.848	0.747
Umich42	top 99.5% of ranked sentences	0.848	0.747
Umich43	top 99% of ranked sentences	0.847	0.745
Umich44	top 98% of ranked sentences	0.845	0.742
Umich45	top 97% of ranked sentences	0.843	0.740

Figure 5: Task 4 - Runs submitted, F-measure on dev/test data, and official F-measure from NIST

6 Task 4

For the final task, we were given the list of all relevant sentences from all 25 documents in the 50 clusters as well as the list of novel sentences from the first 5 documents in all the clusters. Our goal was then to find the novel sentences in the last 20 documents of each cluster. Therefore, we used the first five sentences of clusters N1-N25 of the 2003 data for training and the first five sentences of clusters N26-N50 for dev/test data.

For this task, the naive baseline is to submit all of the sentences known to be relevant in the last 20 documents of each cluster. Since the F-measure of this baseline was quite high, 0.848, we were unable to beat it using any other method. Therefore, as shown in Table 5, we submitted the naive approach as one of our runs. We experimented with our feature discretization procedure as well as with using different cut-off values on our ranked list of sentences, but as expected, we did not manage to beat the baseline.

7 Conclusions

This was our second year participating in the Novelty track. While last year, we focused on using the MEAD summarizer for detecting new and relevant sentences, this year we experimented with using MEAD sentence features in building a maximum entropy classifier. Our group is still quite new to the area of novelty detection and we have again learned a lot from participating in the TREC evaluation. We found the three new tasks for this year to be challenging, particularly since naive baselines perform rather well in many cases.

For the immediate future, we plan to focus on a number of issues related to this year's novelty evaluation:

- In analyzing our algorithms' performance by cluster, we see that the F-measure can vary widely across clusters. Clearly not all clusters should be treated the same when it comes to novelty detection, and we would like to investigate how the clusters differ and how we might be able to exploit such differences in the future.
- Obviously, by using only six sentence features in our maximum entropy classifiers, we did not take advantage of the fact that this method can easily handle the addition of many features into the model. Therefore, we need to continue to develop sentence features to be used. In particular, we should think about developing features at the cluster or intra-sentence level in addition to sentence level features.
- Finally, we are interested in how the TREC notion of novelty compares to that of others (such as in [2] and [3]). For example, in the TREC novelty track, we consider novelty at the sentence level, whereas the TDT initiative can be thought of as addressing novelty at the document level. We are especially interested in considering how analyzing novelty at different levels of granularity could benefit text summarization and question-answering systems.

8 Acknowledgments

The authors of the Maxent 2.1.0 package we used are Jason Baldridge, Tom Morton, Gann Bierner and Eric Friedman.

This work was partially supported by the National Science Foundation's Information Technology Research program (ITR) under grant IIS-0082884. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the participants, and do not necessarily reflect the views of the National Science Foundation.

References

[1] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. In *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, April 2000.

[2] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal Summaries of News Topics. In *Proceedings of the 24th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 10–18, New Orleans, LA, 2001.

[3] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, February 1998.