ICT's Approaches to HTD and Tracking at TDT2004

Man-Quan Yu[+*] Wei-Hua Luo[+] Zhao-Tao Zhou[+*] Shuo Bai[+]
[+]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080,China
[*]Graduate School of Chinese Academy of Sciences, Beijing, 100039, China
yumanquan@software.ict.ac.cn

ABSTRACT

This is ICT's first year of participation in the TDT evaluation. We participate in two tasks: Hierarchical Topic Detection (HTD) and Tracking. The two systems are both based on vector-space model. We use the method of multi-layered clustering to produce directed acyclic graph (DAG) of topics and improve the performance using the technology of traditional detection task. We only implement a baseline system for the tracking task. Our preprocessor of text help reduce the tracking cost.

1. INTRODUCTION

This paper describes the ICT hierarchical topic detection (HTD) and tracking system designed for the TDT2004 evaluation. The two systems are both based on vector-space model.

The task of HTD took place for the first time in TDT 2004. It is intended to overcome two problematic assumptions that all topics are at the same level of granularity, and the assumption that each story pertains to at most one topic. Unlike traditional detection task, each HTD system must construct a DAG over the designated collection of topics. The layers of the DAG represent increasing granularity, with the root vertex being most general and the leaf vertices being most specific. The task may be treated as retrospective search. In this task, we aimed to try to improve the performance of our dry run system under traditional detection evaluation metric and to validate how much the traditional detection system can benefit the new HTD task.

According to previous evaluations, special usage of name entity, time function and clustering in time order were effective methods of topic Detection. We tried these methods on the TDT4 corpus retrospectively and achieved $(C_{det})_{norm}$ of 0.1578 and 0.1523 for the English and Mandarin corpora. To produce DAG for the HTD task, we changed our system from one layered clustering into multi-layered clustering. The higher the layer, the smaller the threshold. It clusters circularly until the root is produced. So the clusters of the higher layers represent the more general topics.

This year, the TDT5 corpus is much larger than previous TDT corpora. There are 407,505 stories in the corpus. Broken down by language it has 72,910 Arabic stories, 278,109 English stories, and 56,486 Mandarin stories. The large size of corpus has brought great challenges to the clustering algorithm. Furthermore, the HTD task requires more complicated methods than usual to produce more complicated topic structures in essence. So there should not only be good performance but good efficiency in the clustering methods. In experiments, we found that the centroid-based system performs well in all aspects.

In the multilanguage condition, we tried several methods to solve the differences of languages, including smoothing the threshold based on language sources, using language-based central vectors and etc. In the evaluation, we first clustered the English native stories into DAG, and then inserted the Mandarin and Arabic stories into the vertices. Experiments showed that this strategy performs well in the multilanguage condition than our original system.

In topic tracking, we only participate in the primary subtask. Given only one story on a particular topic, the subtask of a tracking system is to process a supplied list of data files, and classify all stories in these files as either on-topic or off-topic. This year, we only changed our detection system to suit the tracking task, no document expansion and adaptive learning were applied. And we only did some work on the preprocessor of text.

2. Hierarchical Topic Detection

2.1 System Overview

In TDT2004, ICT divided HTD into four stages. It includes: text preprocessing, feature selection, batching and clustering in buckets, multi-layered clustering. The architecture of our hierarchical topic detection system is depicted in Fig. 1.

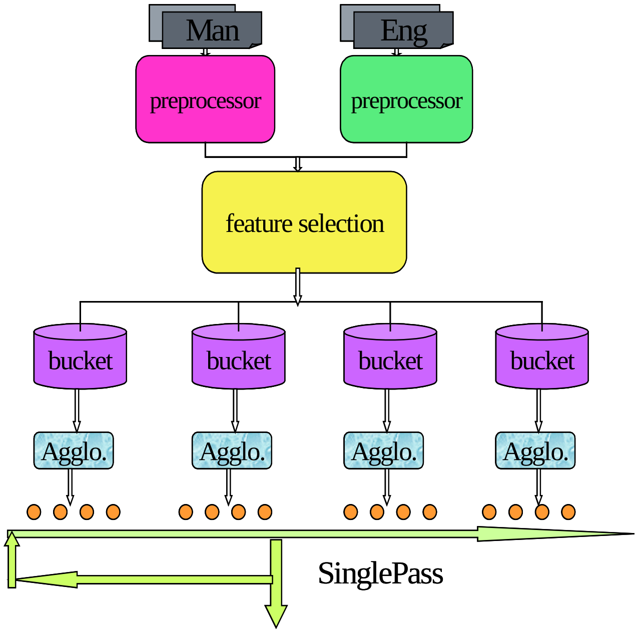

Figure 1: Components of the HTD System

In the preprocessing stage, we process the Mandarin native stories and the English stories separately. For the Mandarin native stories, we applied word segmentation, part-of-speech tagging and name entity identification. For the English stories, we applied tokenization, part-of-speech tagging and morphological analysis. In the stage of feature selection, we first remove stop words and stop part-of-speech. So only noun, adj., name entity and some of special symbols are kept. Then all the remaining words whose value of DF is smaller than 3 or is bigger than N/3 (N is the counts of the whole stories) are removed. In the third stage, we batch the stories into many buckets by time order, and using agglomerative hierarchical clustering method in each bucket to produce micro-clusters. Experiments showed that batching and clustering improves not only efficiency but also performance. Then in the last, multi-layered clustering is used to produce DAG. Our system is rooted in vector-space model (VSM). Based on VSM, we tried different feature weighting and similarity measuring methods.

2.2 Multi-layered Clustering

In HTD task, topics are constructed into DAG. The root vertex of the DAG represents the entire collections. Children of the root represent subsets of stories (which may be overlapping). At each successive layer of the DAG, vertices represent subsets of their parent clusters. Again, each subset may overlap with other subsets.

Generally, one can use different thresholds to produce vertices of different granularities and use the count of the thresholds to control the depth of DAG. Agglomerative and Bisecting methods can all be used. To solve the problem of overlapping, a cluster of the lower layer can be combined into one or more clusters of its higher layer. Our algorithm is an agglomerative method. Firstly, it starts at a certain threshold to cluster in the bottom layer, then the threshold is deceased by a certain distance and clusters again to produce the vertices of the higher layer. This process goes on circularly until the root of DAG is generated. Fig.2 is the flow chart of multi-layered clustering algorithm.

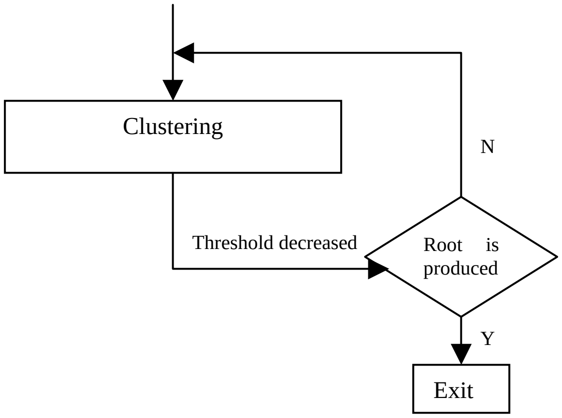

Figure 2: Multi-layered Clustering flow chart

Compared with previous evaluations, the cost of travel is considered into the HTD evaluation metric. The structure of DAG influences $(C_{travel})_{norm}$ in a large degree. Concretely, the depth of DAG lies on the layers of clustering and the width of DAG lies on the counts of clusters of each layer. For the clustering algorithms of each layer, we tried two methods: singlepass and kmeans. The count of results of singlepass is determined by the threshold of similarity and the one of kmeans is mainly determined by the value of k.

Singlepass has been widely used in previous evaluations. In this method, singlelink + time has achieved good performance. However, singlelink has too high time complexity to apply to corpus with large size. We compared the performance and efficiency of singlelink system with centroid-based system using TDT4 Mandarin corpus. The comparison was taken under traditional detection evaluation metric. The results of comparison are showed in Fig. 3 and Fig. 4.

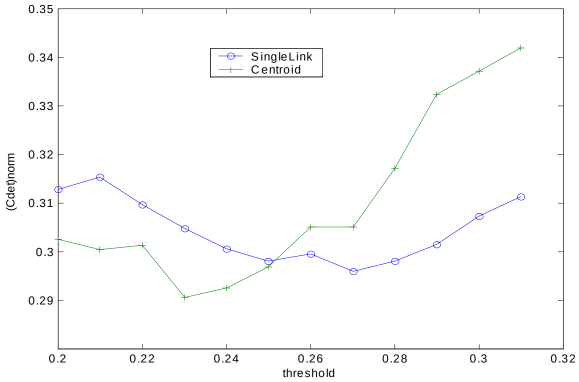

Figure 3: comparison of performance

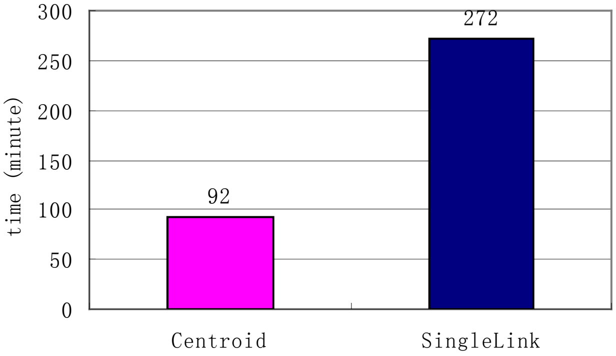

Figure 4: comparison of efficiency

The count of the stories in TDT4 Mandarin corpus is 27,142. From Fig. 3 and Fig. 4, we can see that the performance of centroid-based system is almost the same as singlelink system while the time it consumes is much less. Specially, we achieved better performance using centroid-based system. This is maybe due to other strategies that we used. So in the formal evaluation, we chose centroid-based system.

In addition, we tried kmeans algorithm for the clustering in each layer. We wanted to use the value of k to control the count of clusters in each layer and to make the DAG structure more orderly. But it didn't achieve good result as expected compared with using singlepass in experiments.

2.3 Batching by Time

In order to achieve good performance with HTD, we not only pursued good structure of DAG to minimize $(C_{travel})_{norm,}$ but tried our best to minimize $(C_{det})_{norm}$ of each layer. We did this retrospectively under traditional evaluation metric.

A big difference between TDT and information retrieval is that in TDT the tasks are intimately related to topics and time. To deal with this condition, it is necessary to adjust the traditional IR methods to fit the characteristics of topics. Selecting representative features, time-decayed similarity function and giving priority of clustering in deferral period have been good experience in TDT. Because this year maximum deferral period is not required, we paid attention to the influence of clustering in time batches. In the experiments using TDT4 Mandarin corpus, we made the stories into many batches by time order. Then in each batch, we clustered the stories using agglomerative hierarchical method into micro-clusters. Experiments showed that batching and clustering can improve the performance greatly.

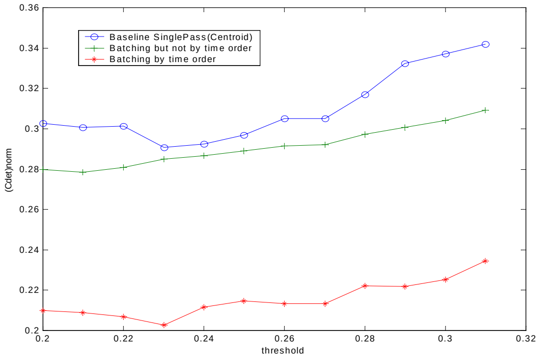

Figure 5: Results of Batching and Clustering

In Fig.5, Curve 1 is the result of the baseline singlepass system using various thresholds. Curve 2 is the result of batching by time order. Here we used time window of 7 days. We can see that the performance was improved nearly 30% than baseline system. Through further experiments, we drew a conclusion that the improvement of performance was mainly due to the priority of merging of the same topics. In the experiment of Curve 3, we still batched the stories and clustered each batch using agglomerative method, but what is different from Curve 2 is that we break up the time order of the stories equably, that is, the time of the stories in each bucket is not adjacent. In this way, we found that the performance improved slightly but was much worse than that of batching by time order. With the increasing of time window, the performance had a tendency to improve but started to decline when a certain value of time window reached. So in the evaluation, we used the strategy of batching by time order firstly before multi-layered clustering was used.

2.4 Feature Weighting

Another important issue is weighting of individual features. We tried several methods of feature weighting. It includes: ltc, InQuery tf·idf, Okapi and Lnu-Ltu weighting scheme. The former three schemes have been fully used in the past evaluations. And the Lnu weighting of a term in a document is defined as

$$\frac{\dfrac{1+\log(tf)}{1+\log(average\_tf)}}{(1.0-slope)*pivot+slope*\#\_of\_unique\_terms} \qquad (1)$$

where slope and pivot are two parameters for pivoted normalization, which can be learned through training. tf is the term frequency in the document. #_of_unique_terms is the number of different terms that occur in the document. Pivoted normalization can be used to modify any normalization function to reduce the gap between the relevance and the retrieval probabilities [13].

In experiments, we found that ltc weighting scheme performed best in our system. And Lnu-Ltu got the same results as ltc. The results of ltc and Inquery tf_idf are depicted in Fig.6.

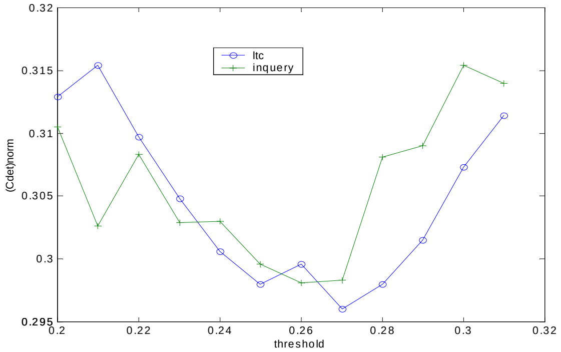

Figure 6: Results of feature weighting

2.5 Strategies for Multilanguage Condition

This year, the multilingual task (English, Mandarin and Arabic) is required of all systems, as usual for topic detection. And the English only task is also required of all systems. The Mandarin only and Arabic only tasks are optional. In the multilanguage condition, the stories of Mandarin and Arabic are translated into English by software of machine translation.

In experiments under traditional detection metric, when we used the monolingual system in multilingual condition, the performance became very poor. The $(C_{det})_{norm}$ is about three times than that of monolingual task. We doubted that this may be caused by the mismatch of words between different languages. So we made experiments with the combination of two languages separately. The experimental result is showed in Table 1.

Language	Eng	Man	Arb
Eng	0.1578	0.1653	0.2964
Man		0.1673	0.5741
Arb			0.6442

Table 1: $(C_{det})_{norm}$ of combinations of different languages

From Table 1, we can see that there was good performance with English and Mandarin stories, no matter if they were combined or not. But when we considered into the translated Arabic stories, the performance declined quickly. Compared with monolingual condition, the performance declined 66% when Arabic stories were combined into English stories and 207% when Arabic stories were combined into Mandarin stories. From this phenomenon, we thought that there was mismatch with words between English, Mandarin and Arabic stories in TDT4 corpus. And this was more serious between Mandarin and Arabic stories.

It is noticeable that the result of the Arabic monolingual condition was the worst.

We tried several methods to solve the mismatches between languages, including smoothing the threshold based on language sources, using language-based central vectors and etc. In the evaluation, we first clustered the English native stories into clusters, and then inserted the Mandarin and Arabic stories into the vertices. Experiments showed that this strategy performs well in the multilanguage condition than our original system. Table 2 and Table 3 are the results of the improved system under detection and HTD evaluation metric. The corpus we used was TDT4 multilingual corpus.

System	$(C_{det})_{norm}$
Original system	0.5234
Improved system	0.3420

Table 2: Results under detection evaluation metric

System	**MinimumCost**
Original system	0.4245
Improved system	0.4012

Table 3: Results under HTD evaluation metric

2.6 Results and Conclusion

In HTD, we participate in the multilingual, English only and Mandarin only subtasks. In each monolingual subtask, we ran our system by just a parameter sweep. In English only subtask, the parameters we used are as follows:

Time batching window: 7
Threshold of agglomerative clustering: 0.35
Threshold of the first layered SinglePass: 0.35
Threshold decreased per layer: 0.05~0.09

And the results of English only subtask are summarized in Table 4.

RUN	$(C_{travel})_{norm}$	$(C_{det})_{norm}$	**MinimumCost**
ICT3a	0.0858	0.1044	0.0981
ICT3b	0.1052	0.1046	0.1048
ICT3c	0.1013	0.1027	0.1022
ICT3d	0.0767	0.0966	0.0898
ICT3e	0.1043	0.0952	0.0983

Table 4: Summary of ICT's English only HTD results

In Table 4, ICT3a is the result using the threshold distance of 0.05 and ICT3e is the one using 0.09. The others are the results using 0.06,0.07 and 0.08 separately.

From Table 4, we can see that the performance does not change seriously when using different threshold distance. The results using 0.06 and 0.07 are the worst no matter in the evaluation or in our experiments with TDT4 corpus. And the DAG structures of the five runs are showed in Table 5.

Depth	**Width**				
	ICT3a	**ICT3b**	**ICT3c**	**ICT3d**	**ICT3e**
0	1	1	1	1	1
1	2	221	2	60	889
2	321	2,723	625	2,509	10,621
3	2,119	10,928	6,031	14,609	31,514
4	7,518	23,578	18,987	33,939	57,670
5	16,774	38,852	36,418	57,670	87,472
6	28,342	57,670	57,670	87,472	278,109
7	41,217	87,472	87,472	278,109	
8	57,670	278,109	278,109		
9	87,472				
10	278,109				

Table 5: DAG structures of ICT's English only HTD results

The results of Mandarin only and Multilingual subtask are summarized in Table 6.

RUN	$(C_{travel})_{norm}$	$(C_{det})_{norm}$	**MinimumCost**
ICT_Mandarin	0.3472	0.0900	0.1774
ICT_Multilingual	0.0934	0.1212	0.1118

Table 6: Summary of ICT's Mandarin only and Multilingual HTD results

Apparently, the MinimumCost of our Mandarin only result is much worse than that of English only or multilingual results. But when we considered the formula of normalization of travel costs, we could find that the performance of Mandarin only system is almost the same as that of English only system. Normalized travel costs are computed as follows:

$(C_{travel})_{norm} = C_{travel}$ / (CBRANCH * MAXVTS * NSTORIES / AVESPT) + CTITLE

In this formula, CBRANCH=2, CTITLE=1, MAXVTS=3, NSTORIES is the total number of stories in the test set and AVESPT = 88. The meaning of AVESPT is the number of stories per topic corresponding to the average observed in development data. In different language conditions, the number of stories per topic is different. Because the number of stories per topic in Mandarin only condition is much less than that in English only or multilingual conditions, the $(C_{travel})_{norm}$ of Mandarin only system is larger. This caused the larger MinimumCost of Mandarin only system.

3. Topic Tracking

3.1 System Overview

In topic tracking, we only participate in the primary subtask. Given only one story on a particular topic, the subtask of a tracking system is to process a supplied list of data files, and classify all stories in these files as either on-topic or off-topic. This year, we only changed our detection system to suit the tracking task, no document expansion and adaptive learning were applied.

For the method of term weighting, we use a standard version (ltc) of the TF_IDF scheme:

$$Wt(ti) = \log(TF(ti,d)+1)*(\log((N/DF(ti,d))+1)$$

where TF(ti,d) is the weight of term ti in document d, DF(ti,d) is the number of training documents where ti occurs and N is the size of the training corpus used to compute the IDF.

The similarity between the topic profile vector and a document vector is calculated using the well-known cosine measure:

$$\cos(\theta) = \frac{u.v}{|u|.|v|}$$

In the preprocessing stage of English, we mended rule-based POS tagger written by Eric Brill [12]. Besides some bugs in memory, Eric's system cannot directly be employed on original lines but token sequences. And we built a powerful morphological analyzer based on WordNet Codes with an English dictionary. In the preprocessing of Mandarin, we used our HMM based system *ICTCLAS*.

3.2 Results and Discussion

The result of the runs of our system is shown in Table 7. Statistics displayed are topic weighted and macro-averaged.

Run	$(C_{det})_{norm}$	P_{miss}	P_{false}
ICT1_trk	0.5669	0.5460	0.0043

Table 7: Summary of ICT's Primary Tracking result

Table 7 shows that miss rates of our system are high, which results in poor $(C_{det})_{norm.}$ We estimate that this is due to not using document expansion and adaptive learning. And the threshold we used is also set experientially. The trade-off curve of our primary evaluation run is shown in Fig.7.

4. Conclusion and Future Work

In the evaluation of HTD this year, we aimed to achieve good results under traditional detection evaluation metric, and then use it to the new task of HTD. The results showed that a good detection system will help a lot in HTD when considered with DAG structures and the method of multi-layered clustering is effective to produce DAG. And it will improve performance using the methods based upon the characteristics of topics. We believe there are many better methods of producing good topic structures to be explored.

For topic tracking, there is much work for us to do. The most promising ones are to apply document expansion and adaptive learning in our baseline system. We also plan additional efficiency work.

5. ACKNOWLEDGMENTS

The authors would like to thank Hua-Ping Zhang for his intensive help with text preprocessing and thank Hong-Bo Xu for his fruitful discussions and useful advice.

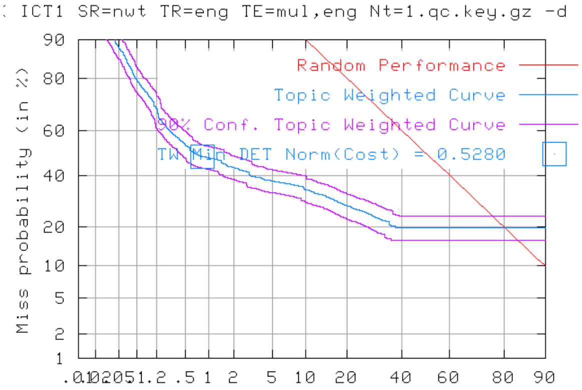

Figure 7: Curve of ICT's Primary Tracking Result

REFERENCES

[1] "The 2004 Topic Detection and Tracking (TDT2004) Task Definition and Evaluation Plan", version 1.0, 5 August 2004.

[2] James Allan (ed.), Topic Detection and Tracking: Event-based Information Organization, Kluwer Academic Publishers, 2002

[3] James Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In Proceedings of SIGIR-98, pages 37{45, Melbourne, Australia, 1998.

[4] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, Topic Detection and Tracking Pilot Study: Final Report, In Proceedings of *the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194-218, San Francisco, CA, 1998, Morgan Kaufmann Publishers, Inc.

[5] Wayne C., Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, *Language Resources and Evaluation Conference* (*LREC*), pages 1487-1494. 2000

[6] Ying-Ju Chen and Hsin-His Chen, NLP and IR Approaches to Monolingual and Multilingual Link Detection, In *Proceedings of the 19th International Conference on Computational Linguistics* (*COLING 2002*)

[7] Yiming Yang, Jaime Carbonell, Ralf Brown, Thomas Pierce, Brian T. Archibald, Xin Liu. Learning Approaches for Detecting and Tracking News Events, *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, Vol. 14(4), pp 32-43, July/August 1999.

[8] Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer, Improving Text Categorization Methods for Event Tracking, In Proceedings of *the 23rd International Conference on Research and Development in Information Retrieval* (*SIGIR-2000*), 2000, pages 65-72

[9] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, The DET Curve in

Assessment of Detection Task Performance, In Proceedings of Eurospeech 1997, pages 1895—1898

[10] R. Willett. Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management*., 25(5):577-597, 1988.

[11] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR 92, Pages 318 – 329,1992.

[12] E. Brill. Some advances in rule-based part of speech tagging. In proceedings of *the twelfth National Conference on Artificial Intelligence* (*AAAI-94*), Seattle, Wa., 1994.

[13] A. Singhal, C. Buckley. and M. Mitra. Pivoted Document Length Normalization. ACM SIGIR, 1996. *http://citeseer.ist.psu.edu/singhal96pivoted.html*