Information Extraction from News Video using Global Rule Induction Technique

[1]Lekha Chaisorn and [2]Tat-Seng Chua

[1]Media Semantics Department, Media Division,
Institute for Infocomm Research (I^2R), Singapore 119613
clekha@i2r.a-star.edu.sg
[2] School of Computing, National University of Singapore,
Singapore 117543
chuats@comp.nus.edu.sg

ABSTRACT

Global rule induction technique has been successfully used in information extraction (IE) from text documents. In this paper, we employ global rule induction technique to perform information extraction from news video documents. We divide our framework into two levels: shot; and story levels. We use a hybrid algorithm to classify each input video shot into one of the predefined genre types and employ the global rule induction technique to extract story boundary information from a sequence of the classified shots. We evaluated our rule induction based system on ~120-hours of news video provided by the TRECVID. And the results show that we could achieve accuracy in F_1-value of over 75%.

1. INTRODUCTION

A two-level multi-modal framework for story segmentation in news video based on Hidden Markov Models (HMM) was proposed in Chaisorn et al. in [5]. The framework composes of two levels: shot and story levels. It employed a hybrid approach and the HMM for the analysis at shot and story levels respectively. The system was evaluated on the TRECVID data set [3] and it achieved the best performing system in the evaluations under story segmentation task [3][6]. But the disadvantages of the system is computational expensive and not easy to scale. Because large amount of news video is rapidly increasing, employing the HMM framework for the analysis will greatly affect the system performance.

Though several methods on story segmentation in news video were proposed, their systems have some limitations. For examples, works in [1] and [2] employed heuristic approach. Their systems yielded good results on their test data. However, Their methods rely heavily on the features from news transcript. Their systems performance will be affected if news transcript is not available. Recent works reported in TRECVID workshop [3][4] used multi-modal features and employed mostly a machine-learning based approach. We know that a framework based on machine-learning approach often encounters data sparseness problem due to insufficient training data. In order to reduce errors from the data sparseness problem, one way is to divide the framework into multiple levels as the concept in NLP research [9] that analyzes text documents in word, phrase and sentence levels. This multiple level concept was also successfully adopted in [5], [6] and [7].

In this research, we employ a global rule induction technique using a system called GRID (Global Rule Induction for Text Documents) [10]. We extend GRID to incorporate the ability to extract story boundary information from video documents based on the two-level framework. The reasons for choosing global rule induction technique are: (1) we observed that when employing the HMM for the analysis at the story level, there are embedded pattern rules in the output; (2) we want to demonstrate the generality and quality of the proposed two-level framework; and (3) to obtain a system that offers less computational cost and less complexity to cope with news story segmentation and indexing in real world problems.

2. THE SYSTEM FRAMEWORK

Briefly, the system composes of shot and story levels as shown in Figure 1. At the lower level, our shot tagger process assigns each of the input shots a unique tag_ID. Each tag_ID represents one of the predefined shot categories. At the higher level, the story extractor process extracts pattern rules (that constitute story boundaries) from a sequence of the tagged shots. Details of the system framework are discussed in the following sections.

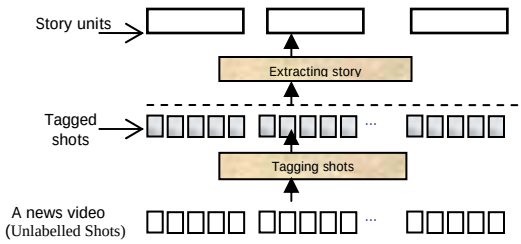

Figure 1: Overview of the System Framework

2.1 Selection of Shot Categories

Most news video has similar structure. It covers news reports on regional and world news, weather, business, sports, commercials, etc. Thus, to reflect this structure and semantics of the shots' contents, we select the following set of shot categories: *Intro/highlight, LEDS, Anchor, 2Anchor, People, Speech/Interview, Live-reporting, Sports, Text-scene, Commercials, Weather, Finance, Special, Program logos* (such as *TOP s*tory, *HEALTH*, *SPORT*, and *PLAY* found in the TRECVID data set). From these categories, we group them into three clusters based on their characteristics. These clusters are: *commercial/heuristic* (CM); *visual similarity-based* (VS)*; and machine-learning-based* (ML). The first cluster contains all commercial shots. The *visual similarity-based* cluster is composed of shots that are visually similar either within a broadcast station or within the same news video. The last cluster, ML cluster, contains the remaining shot categories that cannot be explained using any structures.

The selected categories were demonstrated in [6] and [7] to be useful for story segmentation process. A subset of these categories was also successfully utilized in [8].

2.2 Selection of Features

We select a set of prominent features as given in Table 1. These are the features that are useful to discriminate one shot category from the others. For example, audio class (cheering noise and motion activity are dominant features for "Sports" category, background color and face features are dominant features for "Anchor" and "2Anchor" categories, etc.

Type	Features; Description or format
Low-level	***Color Histogram*;** 176 Luv color histogram
Temporal-level	**-Scene change**;** 1-changed, 0 -unchanged ***-Audio class;*** Speech, music, noise, speech+music, peech+noise, and silence; ***-Motion activity; (***Low, medium, high, or no motion; ***-Shot duration;*** Short, medium, long
Mid-level	***-Face;*** number of faces ***-Shot type;*** Closed-up, medium, long, unknown; ***-Videotexts;*** Lines of text & 1 - center; 0 - /non-centered; **-Cue phrases**;** 1 – presence; 0 - not presence;

Note: Scene change and cue phrase will only be used in story level analysis

Table 1: A set of features selected for the analysis.

The analysis and more details of the selected features can be referred in [6] and [7].

3. TAGGING/MAPPING SHOT

We employ a hybrid approach to map each of the input shots into one of the predefined categories as discussed in *Section 2.1*. Figure 2 illustrates the shot tagging process. Details of the algorithm can be referred in [6] and [7].

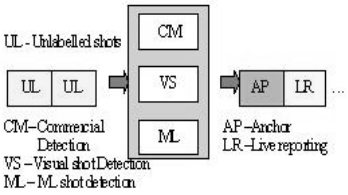

Figure 2: The Shot Tagger contains 3 main sub-processes: CM, VS and ML. It takes unlabelled shots (UL) and classifies them into the predefined categories.

The performance (in F_1-value) of our shot tagging/mapping algorithm testing on a subset of the TRECVID training set [3] is about 90% [6][7].

4. STORY BOUNDARY EXTRACTION USING GLOBAL RULE INDUCTION TECHNIQUE

4.1 Introduction to Global Rule Induction

Global rule induction is an algorithm to induce a set of rules from examples in a form that is similar to if-then-else syntax. Here, we employ GRID system developed by [10] to extract information (rules) from the training data. It emphasizes the use of global feature distribution in all of the training examples in order to make better decision on rule induction. For every training instance, GRID generates a context feature vector centered around the tagged slot (TS). The context feature vector is of the general form:

$$<c_{-k}> \ldots <c_{-2}> <c_{-1}> \; TS \; <c_{+1}> <c_{+2}> \ldots <c_{+k}> \quad (1)$$

$$[<(f_{-k}^{1})..(f_{-k}^{m})>..<(f_{-1}^{1})..(f_{-1}^{m})> \; TS \; <(f_{1}^{1})..(f_{1}^{m})>..<(f_{k}^{1})..(f_{k}^{m})>] \quad (2)$$

Here $<c_i>$, $\{i = -k \text{ to } +k\}$, is the feature vector which represents the context unit of the tagged slot *TS*, and *k* is the number of context units considered. Let *m* be the number of features for each element of the slot. The context feature vector for a single tagged instance can therefore be represented as in Eq (2). By arranging all the instances as in Eq (2), we can obtain a global context feature representation for the training documents as shown in Figure 3.

$$\begin{array}{l} \text{inst.1:}<(f_{-k}^{1}),(f_{-k}^{2})\ldots,(f_{-k}^{m})>..TS..<(f_{+k}^{1}),(f_{+k}^{2})\ldots,(f_{+k}^{m})> \\ \text{inst.2:}<(f_{-k}^{1}),(f_{-k}^{2})\ldots,(f_{-k}^{m})>..TS..<(f_{+k}^{1}),(f_{+k}^{2})\ldots,(f_{+k}^{m})> \\ \quad . \quad . \quad . \quad \ldots \quad . \quad \ldots \quad . \quad \ldots \quad . \\ \text{inst.h:}<(f_{-k}^{1}),(f_{-k}^{2})\ldots,(f_{-k}^{m})>..TS..<(f_{+k}^{1}),(f_{+k}^{2})\ldots,(f_{+k}^{m})> \\ \hline \quad e_{-k}^{1} \quad e_{-k}^{2} \quad e_{-k}^{m} \quad\quad e_{+k}^{1} \quad\quad \ldots \quad\quad e_{+k}^{m} \end{array}$$

Figure 3: Global distribution of *h* instances. The occurrences of the common element features at position *g* are cumulated as e_g^i.

Each extracted rule is constraint using *Laplacian* measure: *Laplacian* (r) = (M +1)/(N+1). Where r is the current extracted rule; M - the number of negative training instances (errors) covered by the rule; and N - the number of positive and negative instances covered by the rule.

In order to understand how GRID learns pattern rules from text documents, we present an example (taken from [10]) as shown in Table 2. In this example, only a subset of feature representations and context elements is presented. And, the semantic slot <stime> which indicates the "starting time" of a seminar announcement is to be extracted.

The example contains 5 positive instances where the desired slots <stime> are tagged at context position 0. By examining the feature frequency for the context elements at every position around the tagged slot, the tagged slot NP_Time (the feature at context position 0) appears most frequently (it occurs 5 times). Thus, it has the highest coverage in the training example pool. This feature is then selected. The generated rule is:

"NP_Time → NP_Time is starting time"

This rule, however, does not satisfy the *Laplacian* measure. Thus, more contextual information must be included to constraint the rule. From the table, the token ":" at the 1st left context position appears 3 times for all positive instances with <stime> in slot 0, it is therefore selected next, and the rule is now constrained as:

": NP_Time → NP_Time is starting time"

Training instances	context position		
	-2	-1	0
inst 1	Time	:	<stime> 3:30 PM </stime>
inst 2	Time	:	<stime> 2 p.m. </stime>
inst 3	Time	:	<stime> 4 p.m </stime>
inst 4	start	at	<stime> 10 am </stime>
inst 5	begin	from	<stime> 11:30 AM </stime>

Table 2: An example for extracting slot "starting time"

For the CMU seminar announcement corpus, this rule is sufficient to satisfy the *Laplacian* measure, and thus we obtained "Rule 1" (see below). Once this rule is obtained, the 3 instances are removed from the positive training example pool. The above process is iterated on the remaining two positive examples and finally we obtain another two rules ("Rule 2" and "Rule 3"). Thus, all the generated rules that satisfied the *Laplacian* measure are:

Rule 1: ": NP_Time → NP_Time is starting time"
Rule 2: "**start at** NP_Time → NP_Time is starting time"
Rule 3: "**begin from** NP_Time → NP_Time is starting time"

During testing, if any of these rules applies, the tags <stime> and </stime> will be inserted beside the NP_Time's boundaries. Further details of GRID can be found in [10].

4.2 Extension of GRID to Story Boundary Extraction

We extend GRID system to incorporate the ability to extract the information in news video. To accomplish story boundary extraction task, we represent each shot using the following features: (a) its tagged category represented by unique ID (obtained from the shot tagging process), (b) scene/location change, (change or unchanged), and (c) cue-phrase at the beginning of story, (present of cue-phrase or no cue-phrase).

A simple example to illustrate how GRID learns pattern extraction rules from news video data is given in Table 3. In the table, there are 5 positive instances selected from our training data. For simplicity, we give an example when using only one feature (shot category) per context unit. The desired slots are tagged as "<>", a dummy slot to indicate story boundary.

Instances	Context position				
	LH_{-2}	LH_{-1}	C_0	RH_{+1}	RH_{+2}
1	SP	LR	<>	LEDS	LR
2	sport	LR	<>	LEDS	sport
3	LR	LR	<>	LEDS	sport
4	sport	sport	<>	LEDS	LR
5	sport	TS	<>	LEDS	SP

Note: SP – Speech; TS – text-scene; LR – Live-reporting; LH – left hand context, RH – right hand context, and C_0 or <> tagged slot (or story boundary)

Table 3: An example for extracting slot "<>" in news video.

We can see from the table that at RH_{+1} position, "LEDS" appears most frequently and thus has the highest coverage in the training example pool. This feature is then selected, and the generated rule is:

"<> LEDS → <> is story boundary"

This rule, however, does not satisfy the *Laplacian* measure as we found that there are many LEDS shots in the negative training instances that are not the successor of story breaks. So the rule has to be constrained. We see that at LH_{-1} position, the token "LR" appears 3 times, and is therefore selected next. The rule is now constrained as:

"LR <> LEDS → <> is story boundary"

In our corpus this rule is sufficient to meet the *Laplacian* measure, and thus the first rule found is:

Rule 1: "LR <> LEDS → <> is story boundary

Once we obtain this rule, we remove the 3 instances from the positive training example pool. The algorithm then iterates the above process on the remaining two positive examples and finally obtains another two rules as follows:

Rule 2: "<sport> <> <LEDS> → <> is story boundary"
Rule 3: "<TS> <> <LEDS> → <> is story boundary".

After all rules are extracted and no more instances in the positive training pool, this set of rules will be used to match the patterns in the testing instances.

5. EXPERIMENTAL RESULTS

The training and testing data sets provided by the TRECVID [3] contains ~120-hours of CNN and ABC news video. We used ~60 hours for training and the rest for testing. We set up several experiments to see the effectiveness of the selected features and the contextual unit length *k*. The results are presented in the following sections.

5.1 Effect of contextual length *k* on different feature sets

In order to investigate the effect of employing different context length on the performance of story segmentation, we conduct a series of experiments using different context length *k*. For each k, the experiments were carried out on the four feature sets (Shot category; Shot category + Scene change; Shot category + cue-phrase; and combination of all of them). The best result for each k is presented in Table 4 (a). The detailed results from the best *k* (*k* =2) are presented in Table 3 (b).

k	F_1-value
1	65.04
2	**75.05**
3	72.63
4	62.25
5	62.26

(a)

Features	F_1-value
Sc	**75.05**
Sc + Cu	62.3
Sc + Cc	59.6
Sc + Cu +Cc	65.0

Sc – Shot category; Cu - Cue phrase; Cc – Scene change

(b)

Table 4: (a) presents the results when using different *k* values. The best result is obtained when *k* equals to *2*. When *k* increases, the performance drops rapidly. Because more context units caused the learnt rules to be more specific, as a result, there are fewer patterns in the test data that match the rules. (b) The performance when using different feature sets and *k* =2. The best performance (~75%) is obtained when using only *shot category* (Sc) as the feature.

5.2 Evaluation Results

Table 5 presents the best results form GRID on the TRECVID data set [3].

Data	Precision	Recall	F_1
ABC	71.95	84.51	**77.72**
CNN	76.76	68.47	72.38
Average	74.36	76.49	75.05

Table 5: Results from GRID on the TRECVID data set. GRID performs better on ABC news video. This is because, CNN data contains variety of programs and the structure is more dynamic.

The results from the HMM on the same data set as reported in [6] and [7] is about 77.5%. Though the HMM is superior to GRID, GRID offers less computational cost and less complexity. The cost (excluding low level feature extraction) based on the training data set is significantly reduced by about 10 times as compared to the HMM.

Figure 4 shows examples of the detected story units in the test instances. More analysis of the results and the generated rules can be found in [7].

6. CONCLUSION AND FUTURE WORK

We proposed a framework to extract story boundary information in news video. We employed a global rule induction technique using GRID system to perform the extraction task. We tested the generality and quality of our framework on the TRECVID data set [3]. Though the HMM yields slightly better results, GRID offers less expensive and easy to scale.

As we can see, here shot genre information is useful for story boundary extraction. And this information is also useful for retrieval of particular shots from news videos, for example, a query on "weather", "stock" or "sports" shots.

Future work is to test the system on other types of videos such as documentary or movie. And, to extend the system to be able to generate news threads to support news retrieval and tracking task.

Figure 4: Examples of the detected stories. (a) Stories detected by the common rule "<> Anchor shot", cover about ~60% on this data set. Stories in (b) is detected by the rule "<> HEALTH". Stories in (c) and (d), the units on the right, are detected by the rules, "Sport <> PLAY" and "Anchor <> Weather" respectively.

REFERENCES

1. A. Merlino, D. Morey and M. Maybury. *Broadcast news navigation using story segmentation*. Proc. of the fifth ACM international conference on Multimedia**,** Seattle, Washington, United States, 1997.
2. A.G. Hauptmann and M. J. Witbrock. *Story Segmentation and Detection of Commercials in Broadcast News Video*. Proc. of Int'l Conference on Advances in Digital Libraries (ADL), California, USA, 1998.
3. //www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2003
4. //www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2004
5. L. Chaisorn, T.-S Chua and C.-H Lee, "The Segmentation of News Video into Story Units", *Proceedings of IEEE Int'l Conf. on Multimedia and Expo*, Lausanne, Switzerland, August 2002.
6. L. Chaisorn, T.-S. Chua, and C.K Koh, Y-L. Zhao, H.-X. Xu, H-.M. Feng. *Story Segmentation and Classification for News Video*. Proceedings of TRECVID 2003, 17-18 November, Washington D.C., USA.
7. L. Chaisorn. "A Hierarchical Multi-modal Approach to Story Segmentation in News Video". PhD thesis, School of Computing, National University of Singapore, 2005.
8. W. H. Hsu and S.-F. Chang. *Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation*. Intl' Conference on Image and Video Retrieval (CIVR) 2005, Singapore.

9. Dale, H. Moisl, and H. Somers. *Handbook of Natural Language Processing.* Marcel Dekker, New York USA, 2000.

10. J. Xiao, T. -S. Chua and J. Liu. *A Global Rule Induction Approach to Information Extraction.* Proc. of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03), 2003.