3WNews: Who, Where, and When in News Video

Jun Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
juny@cs.cmu.edu

Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
alex@cs.cmu.edu

ABSTRACT

We describe *3WNews* as a novel system for browsing news video by the people (who) and locations (where) appearing in the footage as well as the time (when) of news events. The people names, locations, and time expressions are recognized from video transcript and their ambiguous references are resolved. As a key advantage, *3WNews* distinguishes the people and locations that actually appear in the video from those merely mentioned in the transcript, and uses them as (better) indexes for browsing. It also supports browsing of news video by event time instead of broadcasting time.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Design, Management

Keywords: News video, Browsing, Location, People, Time

1. INTRODUCTION

Building systems for accessing news video is important given its rich content and broad audience. Commercial video search engines, such as Google Video (video.google.com), BlinxTV (blinkx.tv), and Youtube (youtube.com), support text-based search but not browsing, which is closer to people's newspaper reading habit. Research systems allow users to browse news video by broadcasting time [1], visual similarity of the video shots, and the presence/absence of semantic concepts [3]. While functionally powerful, these systems can be inconvenient for ordinary users to browse and find their desired news stories.

News video, after all, is about the activities of people in certain locations and at certain time. In this demonstration, we present *3WNews* system—with *3W* standing for *Who*, *Where*, and *When*—for browsing news video by the people and locations that *appear* in the footage as well as the time of news events. Using a Google Map interface augmented by a timeline and a people list, *3WNews* allows users to quickly locate desired news stories by visually specifying the locations, time, and subject people appearing in the stories. It therefore advocates novel access paradigms that are impossible with existing systems, e.g., *"Browse the news footage about Baghdad, Iraq in the past one week"*, and *"Show the footage on President Bush's visit of Europe in 2005"*. Because *3WNews* identifies the locations and people that actually *appear* in the video, it is superior to similar systems such as [1], which organize news video by *all* the locations and people names found in the transcript.

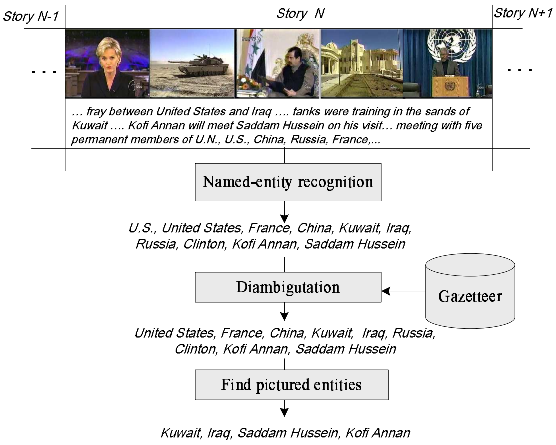
Story N-1
Story N
Story N+1
. . .
. . .
… fray between United States and Iraq …. tanks were training in the sands of Kuwait …. Kofi Annan will meet Saddam Hussein on his visit… meeting with five permanent members of U.N., U.S., China, Russia, France,...
Named-entity recognition
U.S., United States, France, China, Kuwait, Iraq, Russia, Clinton, Kofi Annan, Saddam Hussein
Diambigutation
Gazetteer
United States, France, China, Kuwait, Iraq, Russia, Clinton, Kofi Annan, Saddam Hussein
Find pictured entities
Kuwait, Iraq, Saddam Hussein, Kofi Annan

Figure 1: Extracting pictured locations and person names in a news story at Iraq and Kuwait.

3WNews is based on deep analysis of news video transcript beyond the extraction of bag-of-words, where the transcript comes from either closed-captions or automatic speech recognition. Three problems are addressed: (1) *recognizing* people names, locations, and time expressions (denoted as 3W entities) from the transcript; (2) *disambiguating* the physical references of the recognized 3W entities, e.g., mapping "yesterday" to a specific date, merging synonymous person names and locations; (3) *distinguishing* the people and locations that appear in the video from those merely mentioned in the transcript through syntactic analysis.

The processing behind *3WNews* is exemplified in Figure 1. Recorded news video is first partitioned into *stories* as the basic units for browsing. For each story, we recognize the 3W entities from the corresponding video transcript, resolve the ambiguity as to their physical references, and among them identify the people and locations that likely appear in the video. Then we organize news stories along the people, geographical, and temporal dimension in a novel interface for easy access (see Section 3).

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

2. NAMED ENTITY RECOGNITION AND DISAMBIGUATION

People names, locations, and time expressions are recognized as named entities (NEs) from the video transcript using a named entity detector. Mapping these NEs into unambiguous references of persons, geographical locations, and absolute time stamps requires different techniques.

Person Name: The ambiguity of person names comes from synonymous names, e.g., a person has a canonical (full) name, such as "George Bush", and several variants, such as "President Bush" and "Bush". Merging these synonymous names is important for retrieval and browsing. Within each story, we treat two names identical if one is the *suffix* of the other, such as "Bush" and "George Bush". In addition, we manually create a list of mapping rules between person names and titles, such as "Condoleezza Rice", "Secretary Rice", and "Secretary of the State". This heuristic approach resolves most of the ambiguities on person names.

Location: Locations are more difficult to disambiguate because not only a geographical location may have multiple expressions (synonymity), such as "Great Britain" and "United Kingdom", but one expression can also refer to multiple locations (polysemy), such as "Georgia" as a state in U.S. or as a country. The synonymity problem is solved using a gazetteer, or a dictionary of geographical locations, which contains various names of each location. For the polysemy problem, we examine the context information as the other locations mentioned in proximity of each ambiguous location. For example, if we see "Atlanta" mentioned near "Georgia", "Georgia" is more likely the state of U.S. See [4] for details of disambiguating location names.

Time: The time when a news event took place, or the event time, is different from the broadcasting time of the news. The event time is usually mentioned relative to the broadcasting time, such as "yesterday", "a week ago", and "last Christmas", and we need to transform these relative time expressions into absolute time stamps, such as "2006-05-11". We use the Time Calculus proposed in [2] for anchoring all the relative time expressions.

3. DISTINGUISHING PICTURED PEOPLE AND LOCATIONS

Only part of the person names and locations mentioned in the transcript actually appear in the video, and we call them *pictured* people and locations. In Figure 1, among all the locations mentioned in the news, "Iraq" and "Kuwait" show up in the video while "France", "China", and "Russia" do not. Although all the mentioned people and locations are somewhat related to the news story, those actually appear are more relevant to the video content and are better indexes for searching and browsing the video data. For example, when a user chooses Russia in our map interface, in most cases he/she intends to see news stories happening in Russia, rather than those in which the word "Russia" is mentioned.

We distinguish pictured people and locations from their syntactic roles in the transcript. Intuitively, the way a person's name gets mentioned implies whether the person appears in the video. For example, a name mentioned as the subject of a sentence, e.g, "Bush" in *"President Bush met with the British Prime Minister."* is a stronger indicator of the person's appearance than the name as the modifier of another word, e.g. "Bush" in *"The Bush administration is promoting an industry-friendly plan"*. To capture such clues, we perform syntactic parsing of all the sentences that contain person names, and identify the syntactic role of each name as *subject*, *object*, *modifier*, *prepositional phrase*, etc. We then build a SVM classifier to determine a person's appearance from the syntactic roles of the name in the transcript. Syntactic context is also helpful for determining locations' appearance, but the clues are different from those for person names. For example, a prepositional phrase such as *"in Kuwait"* is a strong indicator of pictured locations. We build a separate classifier for identifying pictured locations.

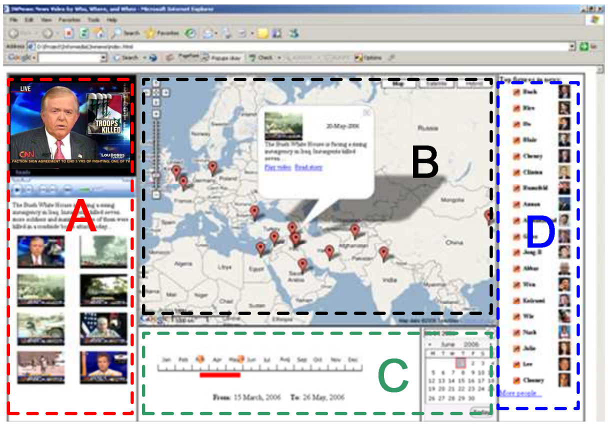

Figure 2: The interface of 3WNews system

4. DEMONSTRATION

We demonstrate the Web browser based interface of *3WNews* system. As shown in Figure 2, the interface consists of 4 panes. Pane B is a map interface built on Google Map API. For each news story, markers (icons) are placed at the predicted locations of that story, and users can click the marker to see the key-frame and summary of that story in a pop-up window. Users can also choose to see the full details of the story and play the video in pane A. Pane C consists of a timeline and a date picker control, which allow users to constrain news stories displayed on the map to those at a specific date or in a certain time range. Pane D contains a list of frequently appearing person names in the news as well as their face pictures. Users can use the checkbox at each name to highlight the markers of the news stories about the corresponding person.

5. REFERENCES

[1] M. G. Christel and et al. Collages as dynamic summaries for news video. In *Proc. of the 10th ACM Int'l Conf. on Multimedia*, pages 561–569, 2002.

[2] B. Han, D. Gates, and L. Levin. From language to time: A temporal expression anchorer. In *Proc. of the 13th Intl Symposium on Temporal Representation and Reasoning*, 2006.

[3] C. G. Snoek and et al. Mediamill: Exploring news video archives based on learned semantics. In *Proc of ACM Multimedia*, pages 225–226, 2005.

[4] J. Yang and A. G. Hauptmann. Annotating news video with locations. In *Proc. of 5rd Int'l Conf. on Image and Video Retrieval*, 2006.