Online Matching of Web Content to Closed Captions in IntoNow

Carlos Castillo*
QCRI
Doha, Qatar
chato@acm.org

Gianmarco De Francisci Morales
Yahoo! Research
Barcelona, Spain
gdfm@yahoo-inc.com

Ajay Shekhawat
Yahoo! Labs
Sunnyvale, USA
ajaysh@yahoo-inc.com

ABSTRACT

IntoNow is a mobile application that provides a second-screen experience to television viewers. IntoNow uses the microphone of the companion device to sample the audio coming from the TV set, and compares it against a database of TV shows in order to identify the program being watched.

The system we demonstrate is activated by IntoNow for specific types of shows. It retrieves information related to the program the user is watching by using closed captions, which are provided by each broadcasting network along the TV signal. It then matches the stream of closed captions in real-time against multiple sources of content. More specifically, during news programs it displays links to online news articles and the profiles of people and organizations in the news, and during music shows it displays links to songs. The matching models are machine-learned from editorial judgments, and tuned to achieve approximately 90% precision.

Categories and Subject Descriptors

[**Information systems**]: Data stream mining; Document filtering;

General Terms

Algorithms

Keywords

Second screen, television companion, news retrieval

1. INTRODUCTION

A *second-screen experience* refers to the use of a companion device, smartphone or tablet, while watching TV. The habit of using a companion device while watching TV is well-established: almost 70% of tablet and smartphone owners use their devices while watching TV [2]. Smart TV sets and second-screen experiences are technologies that have been predicted to have a high impact during 2013 [3]. They promise to revolutionize the way we interact with TV by making it more social and more interactive, and to combine the freedom and richness of web content with the unifying communal experience of watching TV.

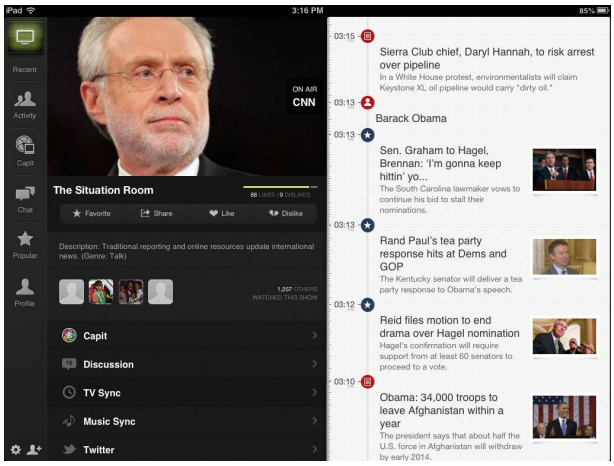

Figure 1: User interface to display related news.

IntoNow is a second-screen application composed of a mobile app and a remote server. The mobile app listens to the audio captured by the device's microphone and transmits a set of fingerprints to the remote server. The server compares them against a database, which is updated in real-time, to identify the show watched by the user. The application then displays the name of the show and links to multiple services including comment forums, screenshots with meme-like messages ("*CapIt*"), and scores and league tables in the case of sport games. Additionally, the application displays *caption-based* information including links to identified news articles (Figure 1) and songs (Figure 2).

The task of automatically selecting information related to closed captions is similar to the one described by Henzinger et al. [1]. Our solution involves web search queries to identify related songs, but not to identify related news. Instead, news matchings are continuously generated on the server side by running several machine-learned models to match the closed captions to a database of web contents.

*Work done while at Yahoo! Research.

Submitted for confidential review.
February 2013.

2. SYSTEM DESCRIPTION

Text pre-processing. The closed captions arrive as timestamped streams of plain text that we process by segmenting into sentences and performing named entity recognition. First, to segment the captions into sentences we use a series of heuristics which include detecting a change of speaker, conventionally signaled by a text marker (“>>”), using the presence of full stops, and using time-based rules. For instance, a pause of several seconds indicates a new sentence.

Second, we extract named entities by using a named entity tagger that performs named entity resolution and ranking [4]. Recognized named entities are associated to weights and represented by URLs of Wikipedia pages.

Example. The input data is similar to this:

```
[1339302650.918]    >> I HOPE YOU GUYS NOW WILL STOP
[1339302653.353]    TALKING ABOUT LEBRON AND THAT HE
[1339302654.555]    DOESN'T PLAY IN BIG GAMES.
[1339302655.856]    HE WAS PRETTY GOOD TONIGHT, WE
[1339302657.558]    CAN GO AHEAD AND PLAY GAME
[1339302659.259]    SEVEN.
[1339302660.000]    >> GAME SEVEN.
[1339302660.467]    WHAT MORE CAN YOU ASK FOR?
[1339302662.169]    >> THIS IS WHAT NBA BASKETBALL
[1339302663.203]    IS ABOUT.
```

The output of our text pre-processing is similar to this:

```
- I hope you guys now will stop talking about
    [entity: LeBron_James] and that he doesn't play
    in big games.
- He was pretty good tonight, we can go ahead and play
    game seven.
- Game seven.
- What more can you ask for?
- This is what [entity: National_Basketball_Association]
    basketball is about.
```

In the output, several lines of closed captions have been joined together to create longer sentences, and named entities such as *LeBron* and *NBA* have been tagged.

News matching. IntoNow obtains the type of each program from an online TV guide. For news matching, we consider programs of type **newscast**, and four genres of news: `general`, `sports`, `entertainment` and `finance`.

We match the processed captions to recent online news articles obtained from hundreds of sources via the Yahoo! News aggregator. Captions are matched in the same genre, e.g., sentences in `sports` are matched to news in the `sports` section of the news aggregator. News in the aggregator that are older than a genre-specific threshold are ignored.

The matching happens in two steps. In the first step, we employ a per-genre classification model trained on thousands of examples labeled by editors. In this model, the two classes are “same story” and “different story”, and each example consists of a sentence and a news story. The features for the classifier are computed from each sentence-story pair by applying the named entity tagger previously described on both elements of the pair, and then by looking at entity co-occurrences. The models are fine-tuned to have high precision of about 90%, with recall between 50% and 60%.

In the second step, recall is improved by aggregating multiple sentence-level matchings that occur in a short time period to form a “qualified matching”. This step allows to leverage the presence of multiple matchings with low confidence close together in time, in order to generate a higher-confidence matching.

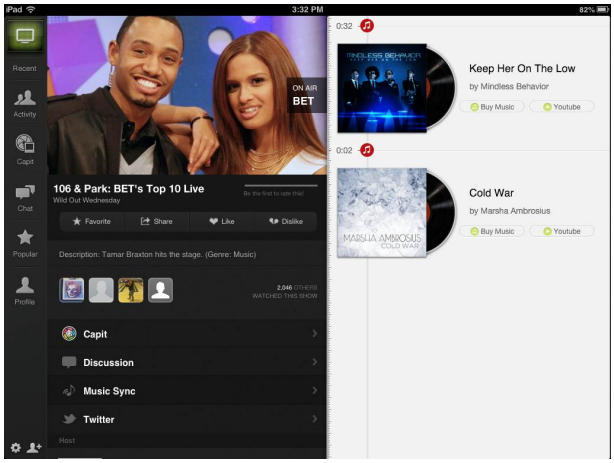

Figure 2: User interface to display detected songs.

In the output, shown in Figure 1, we include links to news articles in Yahoo! News, as well as links to Wikipedia pages for people and organizations mentioned frequently in them.

Song matching. A general convention for TV transcripts of songs is to have a special indicator, a musical note symbol (“♫”), that precedes each line. IntoNow uses this symbol to identify songs. In particular, we use a finite state automaton to decide whether the current note symbol indicates the beginning of a song. The system employs several features to detect state changes, such as the time and the number of captions since the last note symbol.

Once the automaton detects the beginning of a song, it starts accumulating the captions into a buffer. When the buffer reaches a sufficient number of unique non-stop words, we query the Yahoo! Search engine to retrieve pages from the Web. If the result page contains a large number of known lyrics sites, the system extracts the song name and the artist from the title of the pages. We normalize the song names and artists as sites differ in how they list them (e.g., “Romeo” vs “Lil' Romeo”), and count the number of their occurrences. If there is a single song name (and artist) that dominates the distribution of counts, we consider it a match.

We then retrieve matching songs from large online music stores including iTunes and Amazon, and we determine the canonical song name and artist from their results. Finally, we present the user with information about the song, along with links to obtain more information or buy the tracks, as shown in Figure 2.

3. REFERENCES

[1] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-Free news search. *WWWJ*, 8(2):101–126, 2005.

[2] Nielsen. In the U.S., Tablets are TV Buddies while eReaders Make Great Bedfellows. http://blog.nielsen.com/nielsenwire/online_mobile/in-the-u-s-tablets-are-tv-buddies-while-ereaders-make-great-bedfellows, 2011.

[3] H. Taylor. Will social TV and the second screen move beyond advertising in 2013? http://econsultancy.com/uk/blog/11425-will-social-tv-and-the-second-screen-move-beyond-advertising-in-2013, 2013.

[4] Y. Zhou, L. Nie, O. Rouhani-Kalleh, F. Vasile, and S. Gaffney. Resolving surface forms to Wikipedia topics. In *Proc. of COLING*, pages 1335–1343. ACL, 2010.

Requirements

The demonstration will be executed with one laptop and one tablet provided by the authors. Both need Internet access.

The laptop will simulate the role of the TV set and will be used to play pre-recorded YouTube videos of news programs and music shows –or optionally, to access the online stream of a TV network.

The tablet will run the IntoNow application.