Cross-document Event Coreference Resolution based on Cross-media Features

Tongtao Zhang[1], Hongzhi Li[2], Heng Ji[1], Shih-Fu Chang[2]
[1]Computer Science Department, Rensselaer Polytechnic Institute
{zhangt13, jih}@rpi.edu
[2]Department of Computer Science, Columbia University
{hongzhi.li, shih.fu.chang}@columbia.edu

Abstract

In this paper we focus on a new problem of event coreference resolution across television news videos. Based on the observation that the contents from multiple data modalities are complementary, we develop a novel approach to jointly encode effective features from both closed captions and video key frames. Experiment results demonstrate that visual features provided 7.2% absolute F-score gain on state-of-the-art text based event extraction and coreference resolution.

1 Introduction

TV news is the medium that broadcasts events, stories and other information via television. The broadcast is conducted in programs with the name of "**Newscast**". Typically, newscasts require one or several anchors who are introducing stories and coordinating transition among topics, reporters or journalists who are presenting events in the fields and scenes that are captured by cameramen. Similar to newspapers, the same stories are often reported by multiple newscast agents. Moreover, in order to increase the impact on audience, the same stories and events are reported for mutliple times. TV audience passively receives redundant information, and often has difficulty in obtaining clear and useful digest of ongoing events. These properties lead to needs for automatic methods to cluster information and remove redundancy. We propose a new research problem of event coreference resolution across multiple news videos.

To tackle this problem, a good starting point is processing the **C**losed **C**aptions (**CC**) which is accompanying videos in newcasts. The **CC** is either generated by automatic speech recognition (ASR) systems or transcribed by a human stenotype operator who inputs phonetics which are instantly and automatically translated into texts, where events can be extracted. There exist some previous event coreference resolution work such as (Chen and Ji, 2009b; Chen et al., 2009; Lee et al., 2012; Bejan and Harabagiu, 2010). However, they only focused on formally written newswire articles and utilized textual features. Such approaches do not perform well on CC due to (1). the propagated errors from upper stream components (e.g., automatic speech/stenotype recognition and event extraction); (2). the incompleteness of information. Different from written news, newscasts are often limited in time due to fixed TV program schedules, thus, anchors and journalists are trained and expected to organize reports

Figure 1: Similar visual contents improve detection of a coreferential event pair which has a low text-based confidence score.
Closed Captions: "*It 's not clear when it was **killed**.*"; "*Jordan just executed two ISIS prisoners, direct retaliation for the capture of the **killing** Jordanian pilot.*"

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 201–206,
Lisbon, Portugal, 17-21 September 2015.

which are comprehensively informative with complementary visual and CC descriptions within a short time. These two sides have minimal overlapped information while they are inter-dependent. For example, anchors and reporters introduce the background story which are not presented in the videos, and thus the events extracted from CC often lack information about participants.

For example, as shown in Figure 1, these two **Conflict.Attack** event mentions are coreferential. However, in the first event mention, a mistake in Closed Caption ("*he was killed*" → "*it was killed*") makes event extraction and text based coreference systems unable to detect and link "*it*" to the entity of "*Jordanian pilot*". Fortunately, videos often illustrate brief descriptions by vivid visual contents. Moreover, diverse anchors, reporters and TV channels tend to use similar or identical video contents to describe the same story, even though they usually use different words and phrases. Therefore, the challenges in coreference resolution methods based on text information can be addressed by incorporating visual similarity. In this example, the visual similarity between the corresponding video frames is high because both of them show the scene of the Jordanian pilot.

Similar work such as (Kong et al., 2014), (Ramanathan et al., 2014), (Motwani and Mooney, 2012) and (Ramanathan et al., 2013) have explored methods of linking visual materials with texts. However, these methods mainly focus on connecting image concepts with entities in text mentions; and some of them do not clearly distinguish *entity* and *event* in the documents since the definition of visual concepts often require both of them. Moreover, the aforementioned work mainly focuses on improving visual contents recognition by introducing text features while our work will take the opposite route, which takes advantage of visual information to improve event coreference resolution.

In this paper, we propose to jointly incorporate features from both speech (textual) and video (visual) channels for the first time. We also build a newscast crawling system that can automatically accumulate video records and transcribe closed captions. With the crawler, we created a benchmark dataset which is fully annotated with cross-document coreferential events [1].

[1]Dataset can be found at http://www.ee.columbia.edu/dvmm/newDownloads.htm

2 Approach

2.1 Event Extraction

Given unstructured transcribed CC, we extract entities and events and present them in structured forms. We follow the terminologies used in **ACE** (Automatic Content Extraction) (NIST, 2005):

- Entity: an object or set of objects in the world, such as person, organization and facility.
- Entity mention: words or phrases in the texts that mention an entity.
- Event: a specific occurrence involving participants.
- Event trigger: the word that most clearly expresses an event's occurrence.
- Event argument: an entity, or a temporal expression or a value that has a certain role (e.g., Time-Within, Place) in an event.
- Event mention: a sentence (or a text span extent) that mentions an event, including a distinct trigger and arguments involved.

2.2 Text based Event Coreference Resolution

Coreferential events are defined as the same specific occurrence mentioned in different sentences, documents and transcript texts. Coreferential events should happen in the same place and within the same time period, and the entities involved and their roles should be identical. From the perspective of extracted events, each specific attribute and argument from those events should match. However, mentions for the same event may appear in forms of diverse words and phrases; and they do not always cover all arguments or attributes.

To tackle these challenges, we adopt a Maximum Entropy (MaxEnt) model as in (Chen and Ji, 2009b). We consider every pair of event mentions which share the same event type as a candidate and exploit features proposed in (Chen and Ji, 2009b; Chen et al., 2009). Note that the goal in (Chen and Ji, 2009b; Chen et al., 2009) was to resolve event coreference within the same document, whereas our scenario yields to a cross-document/video transcript setting, so we remove some improper and invalid features. We also investigated the approaches by (Lee et al., 2012) and (Bejan and Harabagiu, 2010), but the confidence estimation results from these alternative methods are not reliable. Moreover, the input of event coreference are automatic results from event extraction instead of gold standard, so the noise and errors significantly impact the corefer-

ence performance, especially for unsupervised approaches (Bejan and Harabagiu, 2010). Nevertheless, we still incorporate features from the aforementioned methods. Table 1 shows the features that constitute the input of the MaxEnt model.

2.3 Visual Similarity

Visual content provides useful cues complementary with those used in text-based approach in event coreference resolution. For example, two coreferential events typically show similar or even duplicate scenes, objects, and activities in the visual channel. Coherence of such visual content has been used in grouping multiple video shots into the same video story (Hsu et al., 2003), but it has not been used for event coreference resolution. Recent work in computer vision has demonstrated tremendous progress in large-scale visual content recognition. In this work, we adopt the state-of-the-art techniques (Krizhevsky et al., 2012) and (Simonyan and Zisserman, 2014) that train robust convolutional neural networks (CNN) over millions of web images to detect 20,000 semantic categories defined in ImageNet (Deng et al., 2009) from each image. The 2nd to the last layer features from such deep network can be considered as high-level visual representation that can be used to discriminate various semantic classes (scenes, objects, activity). It has been found effective in computing visual similarity between images, by directly computing the L2 distance of such features or through further metric learning. To compute the similarity between videos associated with two candidate event mentions, we sample multiple frames from each video and aggregate the similarity scores of the few most similar image pairs between the videos. Let $\{f_1^i, f_2^i, ..., , f_l^i\}$ be the key frames sampled from video V_i and $\{f_1^j, f_2^j, ..., , f_l^j\}$ be key frames sampled from video V_j. All the frames are resized to a fixed resolution of 256 x 256 and fed into our pretrained CNN model. We get the high-level visual representation $F_m = FC7(f_m)$ for each frame f_m from the output of the 2nd to the last fully connected layer (FC7) of CNN model. F_m is a 4096 dimension vector. The visual distance of frames f_m and f_n is defined by L_2 distance, which is

$$D_{mn} = ||FC7(f_m^i) - FC7(f_n^j)||_2. \quad (1)$$

The distance of video pair (V_i, V_j) is computed as

$$\bar{D}_{ij} = \frac{1}{k} * \sum_{(f_m, f_n)} D_{mn} \quad (2)$$

, where (f_m, f_n) is the top k of most similar frame pairs. In our experiment, we use $k = 3$. Such aggregation method among the top matches is intended to capture similarity between videos that share only partially overlapped content.

Each news video story typically starts with an introduction by an anchor person followed by news footages showing the visual scenes or activities of the event. Therefore, when computing visual similarity, it's important to exclude the anchor shot and focus on the story-related clips. Anchor frame detection (Hsu et al., 2003) is a well studied problem. In order to detect anchor frames automatically, a face detector is applied to all I-frames of a video. We can obtain the location and size of each detected face. After checking the temporal consistency of the detected faces within each shot, we get a set of candidate anchor faces. The detected face regions are further extended to regions of interest that may include hair and upper body. All the candidate faces detected from the same video are clustered based on their HSV color histogram. It is reasonable to assume that the most frequent face cluster is the one corresponding to the anchor faces. Once the anchor frames are detected, they are excluded and only the non-anchor frames are used to compute the visual similarity between videos associated with event mentions.

2.4 Joint Re-ranking

Using the visual distance calculated from Section 2.3, we can rerank the confidence values from Section 2.2 using the text-based MaxEnt model. We use the following empirical equation to adjust the confidence:

$$W'_{ij} = W_{ij} * e^{-\frac{\bar{D}_{ij}}{\alpha}+1}, \quad (3)$$

where W_{ij} denotes the original coreference confidence between event mentions i and j, D_{ij} denotes the visual distance between the corresponding video frames where the event mentions were spoken and α is a parameter which is used to adjust the impact of visual distance. In the current implementation, we empirically set it as the average of pair-wised visual distances between videos of all event coreference candidates. With this α

Category	Features	Remarks (EM_i: the first event mention, EM_j: the second event mention)
Baseline	type_subtype	pair of event type and subtype in EM_i
	trigger_pair	trigger pair of EM_i and EM_j
	pos_pair	part-of-speech pair of triggers of EM_i and EM_j
	nominal	1 if the trigger of EM_i is nominal
	nom_number	"plural" or "singular" if the trigger of EM_i is nominal
	pronominal	1 if the trigger of EM_i is pronominal
	exact_match	1 if the trigger spelling in EM_i matches that in EM_j
	stem_match	1 if the trigger stem in EM_i matches that in EM_j
	trigger_sim	the semantic similarity scores between triggers of EM_i and EM_j using WordNet(Miller, 1995)
Arguments	argument_match	1 if arguments holding the same roles in both EM_i and EM_j matches
Attributes	mod,pol,gen,ten	four event attributes in EM_i: modality, polarity, genericity and tense
	mod_conflict, pol_conflict, gen_conflict, ten_conflict	1 if the attributes of EM_i and EM_j conflict

Table 1: Features for Event Coreference Resolution

we generally enhance the confidence of event pairs with small visual distances and penalize those with large ones. An alternative way for setting the alpha parameter is through cross validation over separate data partitions.

3 Experiments

3.1 Data and Setting

We establish a system that actively monitors over 100 U.S. major broadcast TV channels such as ABC, CNN and FOX, and crawls newscasts from these channels for more than two years (Li et al., 2013a). With this crawler, we retrieve 100 videos and their correspondent transcribed CC with the topic of "ISIS"[2]. This system also temporally aligns the CC text with the transcribed text from automatic speech recognition following the methods in (Huang et al., 2003). This provides accurate time alignment between the CC text and the video frames. As CC consists of capitalized letters, we apply the true-casing tool from Standford CoreNLP (Manning et al., 2014) on CC. Then we apply a state-of-the-art event extraction system (Li et al., 2013b; Li et al., 2014) to extract event mentions from CC. We asked two human annotators to investigate all event pairs and annotate coreferential pairs as the ground truth. Kappa coefficient for measuring inter-annotator agreement is 74.11%. In order to evaluate our system performance, we rank the confidence scores of all event mention pairs and present the results in Precision vs. Detection Depth curve. Finally we find the video frames corresponding to the event mentions, remove the anchor frames and calculate the visual similarity between the videos. Our final dataset consists of 85 videos, 207 events and 848 event pairs, where 47 pairs are considered coreferential.

We adopt the MaxEnt-based coreference resolution system from (Chen and Ji, 2009b; Chen et al., 2009) as our baseline, and use ACE 2005 English Corpus as the training set for the model. A 5-fold cross-validation is conducted on the training set and the average f-score is 56%. It is lower than results from (Chen and Ji, 2009a) since we remove some features which are not available for the cross-document scenario.

3.2 Results

The peak F-score for the baseline system is 44.23% while our cross-media method boosts it to 51.43%. Figure 2 shows the improvement after incorporating the visual information. We adopt Wilcoxon signed-rank test to determine the significance between the pairs of precision scores at the same depth. The z-ratio is 3.22, which shows the improvement is significant.

For example, the event pair "*So why hasn't U.S. air strikes targeted Kobani within the city*

[2]abbreviation for *Islamic State of Iraq and Syria*

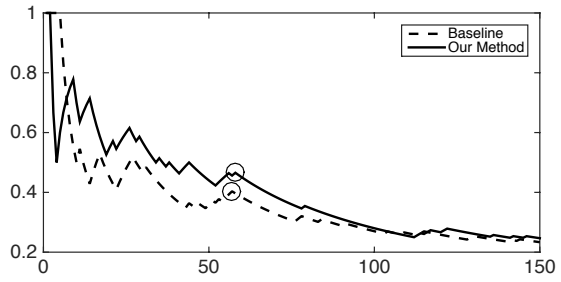

Figure 2: Performance comparison between baseline and our cross-media method on top 150 pairs. Circles indicate the peak F-scores.

limits" and "*Our strikes continue alongside our partners.*" was mistakenly considered coreferential by text features. In fact, the former "strikes" mentions the airstrike and the latter refers to the war or battle, therefore, they are not coreferential. The corresponding video shots demonstrate two different scenes: the former one shows bombing while the latter shows that the president is giving a speech about the strike. Thus the visual distance successfully corrected this error.

3.3 Error Analysis

However, from Figure 2 we can also notice that there are still some errors caused by the visual features. One major error type resides in the negative pairs with both "relatively" high textual coreference confidence scores and "relatively" high visual similarity. From the text side, the event pair contains similar events, for example: "*The Penn(tagon) says coalition air strikes in and around the Syrian city of Kobani have kill hundreds of ISIS fighters but more are streaming in even as the air campaign intensifies.*" and "*Throughout the day, explosions from coalition air strikes sent plums of smoke towering into the sky.*". They talk about two airstrikes during different time periods and are not coreferential, but the baseline system produces a high rank. Our current approach limits the image frames to those overlapped with the speech of an event mention, and in this error, both videos show "battle" scene, yielding a small visual distance. The aforementioned assumption that anchors and journalists tend to use similar videos when describing the same events , which may introduce risk of error caused by similar text event mentions with similar video shots. For such errors, one potential solution is to expand the video frame windows to capture more events and concepts from videos. Expanding the detection range to include visual events in the temporal neighborhood can also differentiate the events.

3.4 Discussion

A systematic way of choosing α in Equation 3 will be useful. One idea is to adapt the α value for different types of events, e.g., we expect some event types are more visually oriented than others and thus use a smaller α value.

We also notice the impact of the errors from the upstream event extraction system. According to (Li et al., 2014) the F-score of event trigger labeling is 65.3%, and event argument labeling is 45%. Missing arguments in events is a main problem, thus the performance on automatically extracted event mentions is significantly worse. About 20 more coreferential pairs could be detected if events and arguments are perfectly extracted.

4 Conclusions and Future Work

In this paper, we improved event coreference resolution on newscast speech by incorporating visual similarity. We also build a crawler that provides a benchmark dataset of videos with aligned closed captions. This system can also help create more datasets to conduct research on video description generation. In the future, we will focus on improving event extraction from texts by introducing more fine-grained cross-media information such as object, concept and event detection results from videos. Moreover, joint detection of events from both sides is our ultimate goal, however, we need to explore the mapping among events from both text and visual sides, and automatic detection of a wide range of objects and events from news video itself is still challenging.

Acknowledgement

This work was supported by the U.S. DARPA DEFT Program No. FA8750-13-2-0041, ARL NS-CTA No. W911NF-09-2-0053, NSF CAREER Award IIS-1523198, AFRL DREAM project, gift awards from IBM, Google, Disney and Bosch. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422.

Zheng Chen and Heng Ji. 2009a. Event coreference resolution: Algorithm, feature impact and evaluation. In *Proceedings of Events in Emerging Text Types (eETTs) Workshop, in conjunction with RANLP, Bulgaria*.

Zheng Chen and Heng Ji. 2009b. Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-4, pages 54–57.

Zheng Chen, Heng Ji, and Robert Haralick. 2009. A pairwise event coreference model, feature impact and evaluation for event coreference resolution. In *Proceedings of the Workshop on Events in Emerging Text Types*, eETTs '09, pages 17–22.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.*, pages 248–255.

Winston Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giridharan Iyengar. 2003. Discovery and fusion of salient multimodal features toward news story segmentation. In *Electronic Imaging 2004*, pages 244–258.

Chih-Wei Huang, Winston Hsu, and Shin-Fu Chang. 2003. Automatic closed caption alignment based on speech recognition transcripts. *Rapport technique, Columbia*.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500.

Hongzhi Li, Brendan Jou, Jospeh G Ellis, Daniel Morozoff, and Shih-Fu Chang. 2013a. News rover: Exploring topical structures and serendipity in heterogeneous multimedia news. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 449–450.

Qi Li, Heng Ji, and Liang Huang. 2013b. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 73–82.

Qi Li, Heng Ji, Yu HONG, and Sujian Li. 2014. Constructing information networks using one single model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1851.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Tanvi S Motwani and Raymond J Mooney. 2012. Improving video activity recognition using object recognition and text mining. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 600–605.

NIST. 2005. The ace 2005 evaluation plan. `http://www.itl.nist.gov/iad/mig/tests/ace/ace05/doc/ace05-evaplan.v3.pdf`.

Vignesh Ramanathan, Percy Liang, and Li Fei-Fei. 2013. Video event understanding using natural language descriptions. In *Proceedings of 2013 IEEE International Conference on Computer Vision*, pages 905–912.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking people in videos with their names using coreference resolution. In *Computer Vision–ECCV 2014*, pages 95–110.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.