Enhancing Multi-lingual Information Extraction via Cross-Media Inference and Fusion

Adam Lee, Marissa Passantino, Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
hengji@cs.qc.cuny.edu

Guojun Qi, Thomas Huang
Department of Electrical and Computer Engineering & Beckman Institute
University of Illinois at Urbana-Champaign
huang@ifp.uiuc.edu

Abstract

We describe a new information fusion approach to integrate facts extracted from cross-media objects (videos and texts) into a coherent common representation including multi-level knowledge (concepts, relations and events). Beyond standard information fusion, we exploited video extraction results and significantly improved text Information Extraction. We further extended our methods to multi-lingual environment (English, Arabic and Chinese) by presenting a case study on cross-lingual comparable corpora acquisition based on video comparison.

1 Introduction

An enormous amount of information is widely available in various data modalities (e.g. speech, text, image and video). For example, a Web news page about "Health Care Reform in America" is composed with texts describing some events (e.g., Final Senate vote for the reform plans, Obama signs the reform agreement), images (e.g., images about various government involvements over decades) and videos/speech (e.g. Obama's speech video about the decisions) containing additional information regarding the real extent of the events or providing evidence corroborating the text part. These cross-media objects exist in redundant and complementary structures, and therefore it is beneficial to fuse information from various data modalities. The goal of our paper is to investigate this task from both mono-lingual and cross-lingual perspectives.

The processing methods of texts and images/videos are typically organized into two separate pipelines. Each pipeline has been studied separately and quite intensively over the past decade. It is critical to move away from single media processing, and instead toward methods that make multiple decisions jointly using cross-media inference. For example, video analysis allows us to find both entities and events in videos, but it's very challenging to specify some fine-grained semantic types such as proper names (e.g. "Obama Barack") and relations among concepts; while the speech embedded and the texts surrounding these videos can significantly enrich such analysis. On the other hand, image/video features can enhance text extraction. For example, entity gender detection from speech recognition output is challenging because of entity mention recognition errors. However, gender detection from corresponding images and videos can achieve above 90% accuracy (Baluja and Rowley, 2006). In this paper, we present a case study on gender detection to demonstrate how text and video extractions can boost each other.

We can further extend the benefit of cross-media inference to cross-lingual information extraction (CLIE). Hakkani-Tur et al. (2007) found that CLIE performed notably worse than monolingual IE, and indicated that a major cause was the low quality of machine translation (MT). Current statistical MT methods require large and manually aligned parallel corpora as input for each language pair of interest. Some recent work (e.g. Munteanu and Marcu, 2005; Ji, 2009) found that MT can benefit from multi-lingual comparable corpora (Cheung and Fung, 2004), but it is time-consuming to identify pairs of comparable texts; especially when there is

Coling 2008: Poster Volume, pages 630–638,
Beijing, August 2010

lack of parallel information such as news release dates and topics. However, the images/videos embedded in the same documents can provide additional clues for similarity computation because they are 'language-independent'. We will show how a video-based comparison approach can reliably build large comparable text corpora for three languages: English, Chinese and Arabic.

2 Baseline Systems

We apply the following state-of-the-art text and video information extraction systems as our baselines. Each system can produce reliable confidence values based on statistical models.

2.1 Video Concept Extraction

The video concept extraction system was developed by IBM for the TREC Video Retrieval Evaluation (TRECVID-2005) (Naphade et al., 2005). This system can extract 2617 concepts defined by TRECVID, such as "Hospital", "Airplane" and "Female-Person". It uses support vector machines to learn the mapping between low level features extracted from visual modality as well as from transcripts and production related meta-features. It also exploits a Correlative Multi-label Learner (Qi et al., 2007), a Multi-Layer Multi-Instance Kernel (Gu et al., 2007) and Label Propagation through Linear Neighborhoods (Wang et al., 2006) to extract all other high-level features. For each classifier, different models are trained on a set of different modalities (e.g., the color moments, wavelet textures, and edge histograms), and the predictions made by these classifiers are combined together with a hierarchical linearly-weighted fusion strategy across different modalities and classifiers.

2.2 Text Information Extraction

We use a state-of-the-art IE system (Ji and Grishman, 2008) developed for the Automatic Content Extraction (ACE) program[1] to process texts and automatic speech recognition output. The pipeline includes name tagging, nominal mention tagging, coreference resolution, time expression extraction and normalization, relation extraction and event extraction. Entities include coreferred persons, geo-political entities (GPE), locations, organizations, facilities, vehicles and weapons; relations include 18 types (e.g. "*a town some 50 miles south of Salzburg*" indicates a located relation.); events include the 33 distinct event types defined in ACE 2005 (e.g. "*Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.*" indicates a "personnel-start" event). Names are identified and classified using an HMM-based name tagger. Nominals are identified using a maximum entropy-based chunker and then semantically classified using statistics from ACE training corpora. Relation extraction and event extraction are also based on maximum entropy models, incorporating diverse lexical, syntactic, semantic and ontological knowledge.

3 Mono-lingual Information Fusion and Inference

3.1 Mono-lingual System Overview

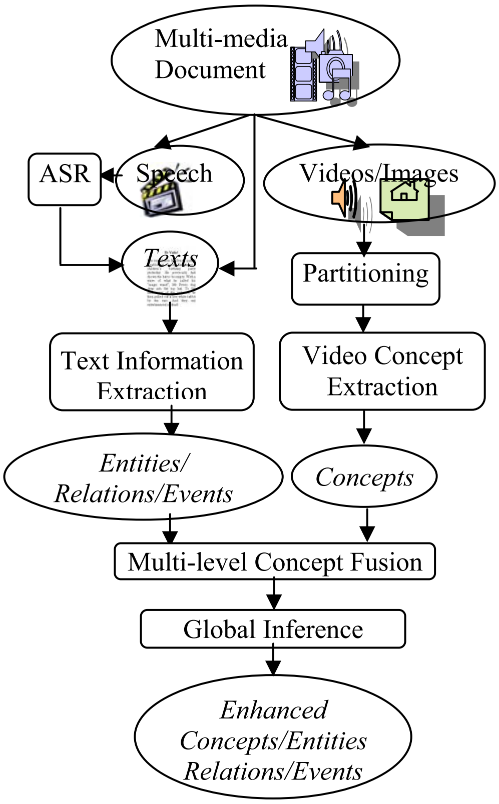

Multi-media Document
ASR
Speech
Videos/Images
Texts
Partitioning
Text Information Extraction
Video Concept Extraction
Entities/ Relations/Events
Concepts
Multi-level Concept Fusion
Global Inference
Enhanced Concepts/Entities Relations/Events

Figure 1. Mono-lingual Cross-Media Information Fusion and Inference Pipeline

Figure 1 depicts the general procedure of our mono-lingual information fusion and inference

[1] http://www.nist.gov/speech/tests/ace/

approach. After we apply two baseline systems to the multi-media documents, we use a novel multi-level concept fusion approach to extract a common knowledge representation across texts and videos (section 3.2), and then apply a global inference approach to enhance fusion results (section 3.3).

3.2 Cross-media Information Fusion

- ***Concept Mapping***

For each input video, we apply automatic speech recognition to obtain background texts. Then we use the baseline IE systems described in section 2 to extract concepts from texts and videos. We construct mappings on the overlapped facts across TRECVID and ACE. For example, "LOC.Water-Body" in ACE is mapped to "Beach, Lakes, Oceans, River, River_Bank" in TRECVID.

Due to different characteristics of video clips and texts, these two tasks have quite different granularities and focus. For example, "PER.Individual" in ACE is an open set including arbitrary names, while TRECVID only covers some famous proper names such as "Hu_Jintao" and "John_Edwards". Geopolitical entities appear very rarely in TRECVID because they are more explicitly presented in background texts. On the other hand, TRECVID defined much more fine-grained nominals than ACE, for example, "FAC.Building-Grounds" in ACE can be divided into 52 possible concept types such as "Conference_Buildings" and "Golf_Course" because they can be more easily detected based on video features. We also notice that TRECVID concepts can include multiple levels of ACE facts, for example "WEA_Shooting" concept can be separated into "weapon" entities and "attack" events in ACE. These different definitions bring challenges to cross-media fusion but also opportunities to exploit complementary facts to refine both pipelines. We manually resolved these issues and obtained 20 fused concept sets.

- ***Time-stamp based Multi-level Projection***

After extracting facts from videos and texts, we conduct information fusion at all possible levels: name, nominal, coreference link, relation or event mention. We rely on the timestamp information associated with video keyframes or shots (sequential keyframes) and background speech to align concepts. During this fusion process, we compare the normalized confidence values produced from two pipelines to resolve the following three types of cases:

- Contradiction – A video fact contradicts a text fact; we only keep the fact with higher confidence.
- Redundancy – A video fact conveys the same content as (or entails, or is entailed by) a text fact; we only keep the unique parts of the facts.
- Complementary – A video fact and a text fact are complementary; we merge these two to form more complete fact sets.

- ***A Common Representation***

In order to effectively extract compact information from large amounts of heterogeneous data, we design an integrated XML format to represent the facts extracted from the above multi-level fusion. We can view this representation as a set of directed "information graphs" $G=\{G_i (V_i, E_i)\}$, where V_i is the collection of concepts from both texts and videos, and E_i is the collection of edges linking one concept to the other, labeled by relation or event attributes. An example is presented in Figure 2. This common representation is applied in both mono-lingual and multi-lingual information fusion tasks described in next sections.

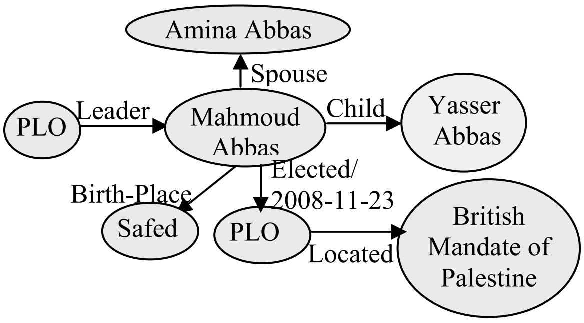


Figure 2. An example for cross-media common fact representation

3.3 Cross-media Information Inference

- ***Uncertainty Problem in Cross-Media Fusion***

However, such a simple merging approach usually leads to unsatisfying results due to uncertainty. Uncertainty in multimedia is induced from noise in the data acquisition procedure (e.g., noise in automatic speech recognition results and low-quality camera surveillance videos) as well as human errors and subjectivity. Unstructured texts, especially those translated from foreign languages, are difficult to interpret. In addition, automatic IE systems for both videos and texts tend to produce errors.

- ***Case Study on Mention Gender Detection***

We employ cross-media inference methods to reduce uncertainty. We will demonstrate this approach on a case study of gender detection for persons. Automatic gender detection is crucial to many natural language processing tasks such as pronoun reference resolution (Bergsma, 2005). Gender detection for last names has proved challenging; Gender for nominals can be highly ambiguous in various contexts. Unfortunately most state-of-the-art approaches discover gender information without considering specific contexts in the document. The results were stored either as a knowledge base with probabilities (e.g. Ji and Lin, 2009) or as a static gazetteer (e.g. census data). Furthermore, speech recognition normally performs poorly on names, which brings more challenges to gender detection for mis-spelled names.

We consider two approaches as our baselines. The first baseline is to discover gender knowledge from Google N-grams using specific lexical patterns (e.g. "[mention] and his/her/its/their") (Ji and Lin, 2009). The other baseline is a gazetteer matching approach based on census data including person names and gender information, as used in typical text IE systems.

We introduce the third method based on male/female concept extraction from associated background videos. These concepts are detected from context-dependent features (e.g. face recognition). If there are multiple persons in one snippet associated with one shot, we propagate gender information to all instances.

We then linearly combine these three methods based on confidence values. For example, the confidence of predicting a name mention n as a male (M) can be computed by combining probabilities $P(n, M, method)$:

$$confidence(n,male) = \lambda_1 * P(n,M,ngram) + \lambda_2 * P(n,M,census) + \lambda_3 * P(n,M,video)$$

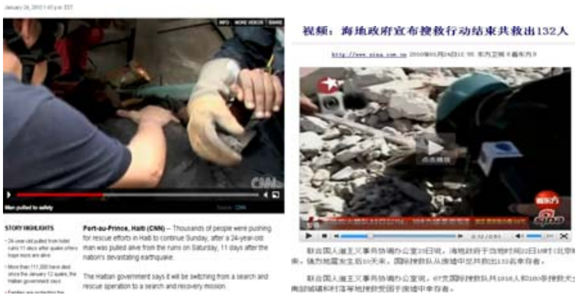
In this paper we used $\lambda_1=0.1$, $\lambda_2=0.1$ *and* $\lambda_3=0.8$ which are optimized from a development set.

4 Cross-lingual Comparable Corpora Acquisition

In this section we extend the information fusion approach to a task of discovering comparable corpora.

4.1 Comparable Documents

Figure 3 presents an example of cross-lingual comparable documents. They are both about the rescue activities for the Haiti earthquake.



Man pulled to safety

STORY HIGHLIGHTS

- 24-year-old pulled from hotel ruins 11 days after quake offers hope more are alive
- More than 111,000 have died since the January 12 quake, the Haitian government says

Port-au-Prince, Haiti (CNN) -- Thousands of people were pushing for rescue efforts in Haiti to continue Sunday, after a 24-year-old man was pulled alive from the ruins on Saturday, 11 days after the nation's devastating earthquake.

The Haitian government says it will be switching from a search and rescue operation to a search and recovery mission.

视频：海地政府宣布搜救行动结束共救出132人

联合国人道主义事务协调办公室23日说，海地政府于当地时间22日[illegible]

Figure 3. An example for cross-lingual multi-media comparable documents

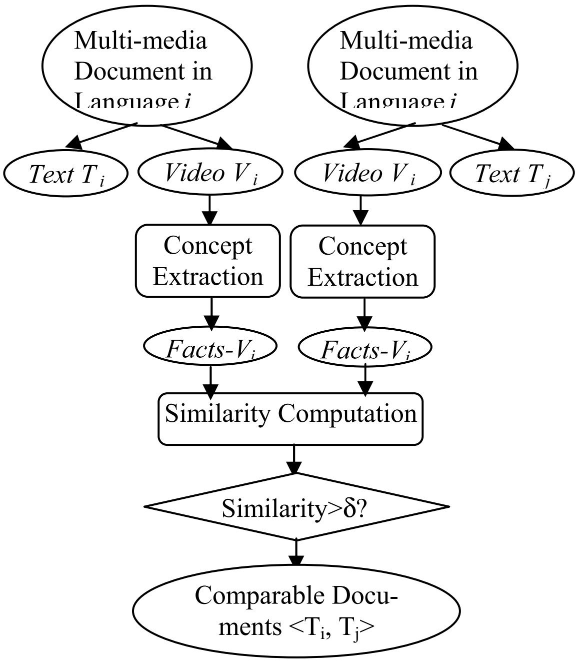


Figure 4. Cross-lingual Comparable Text Corpora Acquisition based on Video Similarity Computation

Traditional text translation based methods tend to miss such pairs due to poor translation quality of informative words (Ji et al., 2009). However, the background videos and images are language-independent and thus can be exploited to identify such comparable documents. This provides a cross-media approach to break language barrier.

4.2 Cross-lingual System Overview

Figure 4 presents the general pipeline of discovering cross-lingual comparable documents based on background video comparison. The detailed video similarity computation method is presented in next section.

4.3 Video Concept Similarity Computation

Most document clustering systems use representations built out of the lexical and syntactic attributes. These attributes may involve string matching, agreement, syntactic distance, and document release dates. Although gains have been made with such methods, there are clearly cases where shallow information will not be sufficient to resolve clustering correctly. Therefore, we should therefore expect a successful document comparison approach to exploit world knowledge, inference, and other forms of semantic information in order to resolve hard cases. For example, if two documents include concepts referring to male-people, earthquake event, rescue activities, and facility-grounds with similar frequency information, we can determine they are likely to be comparable. In this paper we represent each video as a vector of semantic concepts extracted from videos and then use standard vector space model to compute similarity.

Let A=(a_1, …$a_{|\Sigma|}$) and B=(b_1, …$b_{|\Sigma|}$) be such vectors for a pair of videos, then we use cosine similarity to compute similarity:

$$\cos(A,B) = \frac{\sum_{i=1}^{|\Sigma|} a_i b_i}{\sqrt{\sum_{i=1}^{|\Sigma|} a_i^2}\sqrt{\sum_{i=1}^{|\Sigma|} b_i^2}},$$

where $|\Sigma|$ contains all possible concepts. We use traditional TF-IDF (Term Frequency-Inverse Document Frequency) weights for the vector elements a_i and b_i. Let C be a unique concept, V is a video consisting of a series of k shots $V = \{S_1, ..., S_k\}$, then:

$$tf(C,V) = \sum_{i=1}^{k} tf(C,S_i) \Big/ k$$

Let $p(C, S_i)$ denote the probability that C is extracted from S_i, we define two different ways to compute term frequency $tf(C, S_i)$:

(1) $tf(C,S_i) = confidence(C,S_i)$

and

(2) $tf(C,S_i) = \alpha^{confidence(C,S_i)}$

Where *Confidence* (C, S_i) denotes the probability of detecting a concept C in a shot S_i:

$$confidence(C,S_i) = p(C,S_i) \text{ if } p(C,S_i) > \delta, \text{ otherwise 0.}$$

Let: $df(C,S_i) = 1$ if $p(C,S_i) > \delta$, otherwise 0, assuming there are j shots in the entire corpus, we calculate *idf* as follows:

$$idf(C,V) = \log\left(j / \sum_{i=1}^{j} df(C,S_i) \right)$$

5 Experimental Results

This section presents experimental results of all the three tasks described above.

5.1 Data

We used 244 videos from TRECVID 2005 data set as our test set. This data set includes 133,918 keyframes, with corresponding automatic speech recognition and translation results (for foreign languages) provided by LDC.

5.2 Information Fusion Results

Table 1 shows information fusion results for English, Arabic and Chinese on multiple levels. It indicates that video and text extraction pipelines are complementary – almost all of the video concepts are about nominals and events; while text extraction output contains a large amount of names and relations. Therefore the results after information fusion produced much richer knowledge.

Annotation Levels		English	Chinese	Arabic
# of videos		104	84	56
Video	Concept	250880	221898	197233
Text	Name	17350	22154	20057
	Nominal	31528	21852	16253
	Relation	9645	20880	16584
	Event	31132	10348	7148

Table 1. Information Fusion Results

It's also worth noting that the number of concepts extracted from videos is similar across languages, while much fewer events are extracted from Chinese or Arabic because of speech recognition and machine translation errors. We took out 1% of the results to measure accuracy against ground-truth in TRECVID and ACE training data respectively; the mean average precision for video concept extraction is about 33.6%. On English ASR output the text-IE system achieved about 82.7% F-measure on labeling names, 80.5% F-measure on nominals (regardless of ASR errors), 66% on relations and 64% on events.

5.3 Information Inference Results

From the test set, we chose 650 persons (492 males and 158 females) to evaluate gender discovery. For baselines, we used Google n-gram (n=5) corpus Version II including 1.2 billion 5-grams extracted from about 9.7 billion sentences (Lin et al., 2010) and census data including 5,014 person names with gender information.

Since we only have gold-standard gender information on shot-level (corresponding to a snippet in ASR output), we asked a human annotator to associate ground-truth with individual persons. Table 2 presents overall precision (P), recall (R) and F-measure (F).

Methods	P	R	F
Google N-gram	89.1%	70.2%	78.5%
Census	96.2%	19.4%	32.4%
Video Extraction	88.9%	73.8%	80.6%
Combined	89.3%	80.4%	84.6%

Table 2. Gender Discovery Performance

Table 2 shows that video extraction based approach can achieve the highest recall among all three methods. The combined approach achieved statistically significant improvement on recall.

Table 3 presents some examples ("F" for female and "M" for male). We found that most speech name recognition errors are propagated to gender detection in the baseline methods, for example, "Sala Zhang" is mis-spelled in speech recognition output (the correct spelling should be "Sarah Chang") and thus Google N-gram approach mistakenly predicted it as a male. Many rare names such as "Wu Ficzek", "Karami" cannot be predicted by the baselines,

Error analysis on video extraction based approach showed that most errors occur on those shots including multiple people (males and females). In addition, since the data set is from news domain, there were many shots including reporters and target persons at the same time. For example, "Jiang Zemin" was mistakenly associated with a "female" gender because the reporter is a female in that corresponding shot.

5.4 Comparable Corpora Acquisition Results

For comparable corpora acquisition, we measured accuracy for the top 50 document pairs. Due to lack of answer-keys, we asked a bilingual human annotator to judge results manually. The evaluation guideline generally followed the definitions in (Cheung and Fung, 2004). A pair of documents is judged as comparable if they share a certain amount of information (e.g. entities, events and topics).

Without using IDF, for different parameter α and δ in the similarity metrics, the results are summarized in Figure 5. For comparison we present the results for mono-lingual and cross-lingual separately. Figure 5 indicates that as the threshold and normalization values increase, the accuracy generally improves. It's not surprising that mono-lingual results are better than cross-lingual results, because generally more videos with comparable topics are in the same language.

Mention	Google N-gram	Census	Video Extraction	Correct Answer	Context Sentence
Zhang Sala	M: 1 F: 0	-	F: 0.699 M: 0.301	F	World famous meaning violin soloist **Zhang Sala** recently again to Toronto symphony orchestra...
Peter	M: .979 F: 0.021	M: 1	M: 0.699 F: 0.301	M	Iraq, there are in Lebanon Paris pass **Peter** after 10 five Dar exile without peace...
Wu Ficzek	-		M: 0.699 F: 0.301	M	If you want to do a good job indeed **Wu Ficzek**
President	M: .953 F: 0.047	-	M: 0.704 F: 0.296	M	Labor union of Arab heritage publishers **president** to call for the opening of the Arab Book Exhibition.
Jiang Zemin	M: 1 F: 0	-	F: 0.787 M: 0.213	M	It has never stopped the including the former CPC General Secretary Jiang Zemin…
Karami	M: 1 F: 0	-	M: 0.694 F: 0.306	M	all the Gamal Ismail introduced the needs of the Akkar region, referring to the desire on the issue of the President **Karami** to give priority disadvantaged areas

Table 3. Examples for Mention Gender Detection

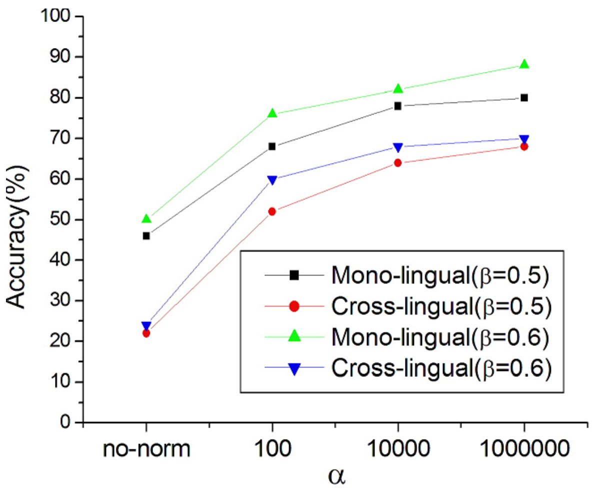


Figure 5. Comparable Corpora Acquisition without IDF

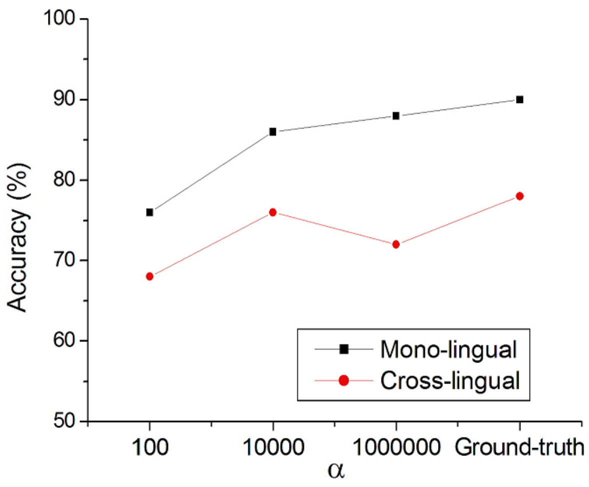


Figure 6. Comparable Corpora Acquisition with IDF (δ=0.6)

We then added IDF to the optimized threshold and obtained results in Figure 6. The accuracy for both languages was further enhanced. We can see that under any conditions our approach can discover comparable documents reliably. In order to measure the impact of concept extraction errors, we also evaluated the results for using ground-truth concepts as shown in Figure 6. Surprisingly it didn't provide much higher accuracy than automatic concept extraction, mainly because the similarity can be captured by some dominant video concepts.

6 Related Work

A large body of prior work has focused on multimedia information retrieval and document classification (e.g. Iria and Magalhaes, 2009). State-of-the-art information fusion approaches can be divided into two groups: formal "top-down" methods from the generic knowledge fusion community and quantitative "bottom-up" techniques from the Semantic Web community (Apriou et al., 2001; Gregoire, 2006). However, very limited research methods have been explored to fuse automatically extracted facts from texts and videos/images. Our idea of conducting information fusion on multiple semantic levels is similar to the kernel method described in (Gu et al., 2007).

Most previous work on cross-media information extraction focused on one single domain (e.g. e-Government (Amato et al., 2010); soccer game (Pazouki and Rahmati, 2009)) and structured/semi-structured texts (e.g. product catalogues (Labsky et al., 2005)). Saggion et al. (2004) described a multimedia extraction approach to create composite index from multiple and multi-lingual sources. We expand the task to the more general news domain including unstructured texts and use cross-media inference to enhance extraction performance.

Some recent work has exploited analysis of associated texts to improve image annotation (e.g. Deschacht and Moens, 2007; Feng and Lapata, 2008). Some recent research demonstrated cross-modal integration can provide significant gains in improving the richness of information. For example, Oviatt et al. (1997) showed that speech and pen-based gestures can provide complementary capabilities because basic subject, verb, and object constituents almost always are spoken, whereas those describing locative information invariably are written or gestured. However, not much work demonstrated an effective method of using video/image annotation to improve text extraction. Our experiments provide some case studies in this new direction. Our work can also be considered as an extension of global background inference (e.g. Ji and Grishman, 2008) to cross-media paradigm.

Extensive research has been done on video clustering. For example, Cheung and Zakhor (2000) used meta-data extracted from textual and hyperlink information to detect similar videos on the web; Magalhaes et al. (2008) described a semantic similarity metric based on key word vectors for multi-media fusion. We extend such video similarity computing approaches to a multilingual environment.

7 Conclusion and Future Work

Traditional Information Extraction (IE) approaches focused on single media (e.g. texts), with very limited use of knowledge from other data modalities in the background. In this paper we propose a new approach to integrate information extracted from videos and texts into a coherent common representation including multi-level knowledge (concepts, relations and events). Beyond standard information fusion, we attempted global inference methods to incorporate video extraction and significantly enhanced the performance of text extraction. Finally, we extend our methods to multi-lingual environment (English, Arabic and Chinese) by presenting a case study on cross-lingual comparable corpora acquisition.

We used a dataset which includes videos and associated speech recognition output (texts), but our approach is applicable to any cases in which texts and videos appear together (from associated texts, captions etc.). The proposed common representation will provide a framework for many byproducts. For example, the monolingual fused information graphs can be used to generate abstractive summaries. Given the fused information we can also visualize the facts from background texts effectively. We are also interested in using video information to discover novel relations and events which are missed in the text IE task.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149, Google, Inc., DARPA GALE Program, CUNY Research Enhancement Program, PSC-CUNY Research Program, Faculty Publication Program and GRTI Program. 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 Army Research Laboratory or 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

Amato, F., Mazzeo, A., Moscato, V. and Picariello, A. 2010. Information Extraction from Multimedia Documents for e-Government Applications. *Information Systems: People, Organizations, Institutions, and Technologies*. pp. 101-108.

Appriou A., A. Ayoun, Benferhat, S., Besnard, P., Cholvy, L., Cooke, R., Cuppens, F., Dubois, D., Fargier, H., Grabisch, M., Kruse, R., Lang, J. Moral, S., Prade, H., Saffiotti, A., Smets, P., Sossai, C. 2001. Fusion: General concepts and characteristics. *International Journal of Intelligent Systems 16(10).*

Baluja, S. and Rowley, H. 2006. Boosting Sex Identification Performance. *International Journal of Computer Vision.*

Bergsma, S. 2005. Automatic Acquisition of Gender Information for Anaphora Resolution. *Proc. Canadian AI 2005.*

Cheung, P. and Fung P. 2004. Sentence Alignment in Parallel, Comparable, and Quasi-comparable Corpora. *Proc. LREC 2004.*

Cheung, S.-C. and Zakhor, A. 2000. Efficient video similarity measurement and search. *Proc. IEEE International Conference on Image Processing.*

Deschacht K. and Moens M. 2007. Text Analysis for Automatic Image Annotation. *Proc. ACL 2007.*

Feng, Y. and Lapata, M. 2008. Automatic Image Annotation Using Auxiliary Text Information. *Proc. ACL 2008.*

Gregoire, E. 2006. An unbiased approach to iterated fusion by weakening. *Information Fusion. 7(1).*

Gu, Z., Mei, T., Hua, X., Tang, J., Wu, X. 2007. Multi-Layer Multi-Instance Kernel for Video Concept Detection. Proc. *ACM Multimedia 2007.*

Hakkani-Tur, D., Ji, H. and Grishman, R. 2007. Using Information Extraction to Improve Cross-lingual Document Retrieval. *Proc. RANLP 2007 Workshop on Multi-Source Multi-lingual Information Extraction and Summarization.*

Iria, J. and Magalhaes, J. 2009. Exploiting Cross-Media Correlations in the Categorization of Multimedia Web Documents. *Proc. CIAM 2009.*

Ji, H. and Grishman, R. 2008. Refining Event Extraction Through Cross-document Inference. *Proc. ACL 2008.*

Ji, H. 2009. Mining Name Translations from Comparable Corpora by Creating Bilingual Information Networks. *Proc. ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora (BUCC 2009): from parallel to non-parallel corpora.*

Ji, H., Grishman, R., Freitag, D., Blume, M., Wang, J., Khadivi, S., Zens, R., and Ney, H. 2009. Name Translation for Distillation. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation.* Springer.

Ji, H. and Lin, D. 2009. Gender and Animacy Knowledge Discovery from Web-Scale N-Grams for Unsupervised Person Mention Detection. *Proc. PACLIC 2009.*

Oviatt, S. L., DeAngeli, A., & Kuhn, K. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. *Proceedings of Conference on Human Factors in Computing Systems (CHI'97)*, 415-422. New York: ACM Press.

Labsky, M., Praks, P., Sv´atek1, V., and Svab, O. 2005. Multimedia Information Extraction from HTML Product Catalogues. *Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence.* pp. 401 – 404.

Lin, D., Church, K., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., Patil, K., Pitler, E., Lathbury, R., Rao, V., Dalwani, K. and Narsale, S. 2010. New Data, Tags and Tools for Web-Scale N-grams. *Proc. LREC 2010.*

Magalhaes, J., Ciravegna, F. and Ruger, S. 2008. Exploring Multimedia in a Keyword Space. *Proc. ACM Multimedia 2008.*

Munteanu, D. S. and Marcu D. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*. Volume 31, Issue 4. pp. 477-504.

Naphade, M. R., Kennedy, L., Kender, J. R., Chang, S.-F., Smith, J. R., Over, P., and Hauptmann, A. A light scale concept ontology for multimedia understanding for TRECVID 2005. *Technical report, IBM, 2005.*

Pazouki, E. and Rahmati, M. 2009. A novel multimedia data mining framework for information extraction of a soccer video stream*.* *Intelligent Data Analysis.* pp. 833-857.

Qi,G.-J., Hua,X.-S., Rui, Y., Tang, J., Mei, T., and Zhang,H.-J. 2007. Correlative Multi-label Video Annotation. *Proc. ACM Multimedia 2007*.

Saggion, H., Cunningham, H., Bontcheva, K., Maynard, D., Hamza, O., and Wilks, Y. 2004. Multimedia indexing through multi-source and multi-language information extraction: the MUMIS project. *Data Knowlege Engineering*, 48, 2, pp. 247-264.

Wang, F. and Zhang, C. 2006. Label propagation through linear neighborhoods. *Proc. ICML 2006.*