Using Abduction for Video-Text Coreference

Jerry R. Hobbs and Rutu Mulkar-Mehta

Information Sciences Institute
University of Southern California
Marina del Rey, California
hobbs@isi.edu,rutu@isi.edu

Abstract. We describe preliminary efforts to use an abductive theorem-prover, called Mini-Tacitus, to interpret the combination of video and audio streams of news broadcasts, aiming at a coherent interpretation of the two together. As a side effect, this process identifies entities and events that occur in both streams, across the modalities. Because vision research are not yet quite up to the task of analyzing the video into the required meaningful predicates, we begin with a hand-encoding of the video stream into a logical form; similarly, we begin with a transcript of the audio stream. We translate the transcript into logical form. We then use knowledge of the structure of events and of the environment to infer the most plausible and coherent interpretation of the sentences, the discourse, and the video, in combination.

1 Abduction

We use abduction to interpret our environment (Hobbs et al., 1993). That is, we try to come up with the best explanation for the observables we encounter, where the explanation consists of things we know and things we assume because they entail the observables. An important part of our environment are the communicative acts of other people, including utterances and displays of various sorts. The best explanation for such acts is usually that they are intentional actions in a plan aimed at achieving some larger goal. The goal of a communicative act is for the hearer to believe or otherwise consider the meaning or content of the communicative act.

Determining the meaning of a string of words is also a matter of abduction, of coming up with the best explanation for what the string of words conveys, or the situation that it describes. This involves figuring out what each of the words conveys and what their order conveys. For the latter, above the level of the sentence, this is a matter of figuring out what "coherence relations", such as causality and exemplification, explain why two segments of discourse are adjacent (Hobbs, 1985a). Within sentences, the best explanation of adjacency is usually the predicate-argument relation; this in fact can be taken as a definition of syntax.

Proceeding in this fashion, we end up with a logical form for the text in which the predicate-argument relations deriving from words are all represented and at

least the fact that adjacent segments are coherently related is represented. Then we must come up with the best explanation, or best abductive proof, for that logical form. We have to construct a scenario that makes sense of the content of the discourse.

Several factors go into what makes an interpretation the *best* interpretation. One of the most important factors is the *economy* of the proof. An interpretation is better if it makes fewer assumptions and has shorter proofs. One way of achieving this is by recognizing the inherent redundancy in discourse. The same underlying ideas are implicit in many different explicit signals in the message, and the same entities are being talked about in different guises. By exploiting this implicit redundancy, we, as a by-product, solve the coreference problems in discourse. Two phrases or other signals describe the same entity, and by recognizing this we solve the coreference problem.

In this paper we describe a very preliminary effort to apply this framework to the problem of video-text coreference, or multimedia interpretation. This work is very similar in approach to that of Espinosa et al. (2007) except that where they use description logic we use unrestricted first-order logic. In this paper we examine one short stretch of a videotaped news broadcast as an example of how this approach would be applied. This data and the goals of the effort are described in Section 2.

In Section 3 we describe the structure of the video data and the structure of the discourse in the audio stream, including the identifications we want to make in each modality and across modalities. In Section 4 we look more closely at several coreference examples the data presents, including examples of video-video, text-video, and text-text coreference. We have implemented an abductive theorem-prover called Mini-Tacitus, and we illustrated the output of the program for the examples we examine.

In Section 4 we describe how the structure and co-reference relations might be recognized dynamically, as the data is fed into the system one phrase and one frame at a time.

2   The Data and the Problem

The data we analyze here is a short stretch from a news broadcast story about the economic decline in the state of Wyoming. People are losing their jobs and have to move to another state for work. The spoken commentary is as follows:

(1) Even those who want to stay are struggling.
(2) Rosey Gallegos and her husband have enjoyed living here in Wyoming for over a decade.
(3) But now, his oil industry job has been moved to Texas and they are moving with it.

While this is being said, the video first shows a woman in her kitchen reaching for and picking up a vase from on top of a cupbord and carrying it over to a

table and putting it in a box. Then it shows a man wheeling some boxes out of a house and loading them into a moving van.

The problem is to identify the entities and events that occur in both the audio and video streams, across the modalities.

Fig. 1. Video Logical Form for the First Three Frames

We took as our starting point the logical forms of the video stream and the audio stream. These were manually encoded. The audio logical forms are flat expressions in first-order logic whose predicates are derived directly from the lexical items (Hobbs, 1985b). The video logical forms consist of static descriptions of frames at half-second intervals, including descriptions in first-order logic of the entities in the frame and the relations between them. **??** shows the video logical forms for the first three frames. The logical form of the text is as follows:

Even those who want to stay are struggling.
$Plural(x11, s11) \wedge want'(e11, x11, e12) \wedge stay'(e12, x11) \wedge struggle'(e13, x11, e14)$
Rosey Gallegos and her husband have enjoyed living here for over a decade.
$Rosey(r1) \wedge Gallegos(r1) \wedge husband(h1, r1) \wedge andn(x21, r1, h1) \wedge poss(r1, h1) \wedge$
$enjoy'(e21, x21, e22) \wedge live'(e22, x21) \wedge perfect(e21) \wedge present(e22) \wedge$
$for(e22, t11) \wedge duration(e22, t11) \wedge over(t11, t12) \wedge decade(t12) \wedge$
$atLocation(e22, l11) \wedge in(l11, w1) \wedge Wyoming(w1)$
But now, his oil industry job has been moved to Texas and they are moving with it.

$$but(e21, e36) \wedge now(e35, u31) \wedge person(x31) \wedge male(x31) \wedge poss(x31, j) \wedge$$
$$job(j) \wedge oil(o) \wedge industry(i) \wedge nn(o, i) \wedge nn(i, j) \wedge move'(e31, z1, j1, w1, t1) \wedge$$
$$Texas(t1) \wedge and'(e35, e31, e36) \wedge move1'(e36, s1, w1, t1) \wedge plural(x31, s31) \wedge$$
$$with'(e37, e36, j1)$$

Both varieties of logical form are first-order logical encodings of the information conveyed in the medium. Thus, the representation scheme is the same, and we will assume we have a set of axioms that link the predicates used in each, if they are not already the same. Our problem is to produce a unified explanation of the two sets of logical forms together.

For the audio stream, we are assuming that speech recognition has produced a correct string of words, and that syntax and compositional semantics have produced a correct logical form. We are aware that these assumptions are overly optimistic, but significant progress is being made in both these areas, and the assumptions have enabled us to focus specifically on the video-text coreference problem.

Similarly, in constructing the logical forms for the video streams, we are positing descriptions of entities at a higher level than current visual processing can produce. For example, we call something a cupboard or a stove despite the fact that visual processing could only say that they are rectilinear objects. One could imagine working from a lower-level vocabulary and using abduction and cross-modal interpretation to determine what objects these are most likely to be, and indeed Shanahan (2005) argues for precisely this sort of approach to perception. But for this initial exercise, we began at the higher level.

3   Discourse and Video Structure

The structure of coherent discourse arises from relations between the eventualities that are the main assertions or primary claims of successive segments of discourse. These "coherence relations" are what the adjacency of the segments are intended to convey, and in finding them, we find the explanation for that adjacency. When a coherence relation is discovered between two segments, the two together constitute a composite segment, whose primary claim is determined from the primary claims of the two constituent segments and the relation between them. In this way, a tree-like structure is built up over the entire discourse (Hobbs, 1985a).

The structure of discourse (1-3) is illustrated in Figure 2. This is easiest to explain from the bottom up.

The two clauses of sentence (3) are related explicitly by the conjunction "and". The two most common interpretations of "and" are similarity and a temporal succession relation—"and then"— that Hobbs (1985a) has called the "occasion" relation. A somewhat less frequent but stronger interpretation of "and" is causality. Here the relation is causality. There is a change in location of the job, one's residence is normally where one's job is, and thus the change in location of the job causes the change in location of their residence. Causal relations often function as explanations, and in explanations the primary claim of
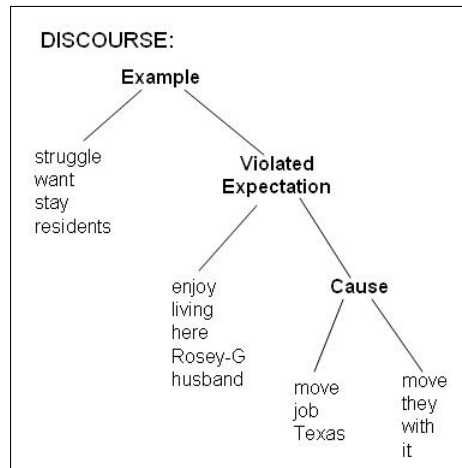
Fig. 2. Graphical Depiction of the Discourse in a Tree-like structure

the composed segment is what is being explained, i.e., the effect. This eventuality, the people's moving, is thus what needs to be linked when we are linking sentence (3) with the preceding discourse. Note that in the course of this reasoning we resolve "it" to "his job".

A weaker but easier-to-discover possible relation between the two clauses is similarity, based on the moving events in each clause. One reason not to favor this interpretation is that the entities that are moving— a job and people—are not terribly similar. In addition, with the similarity relation, the primary claim of the entire segment would be something like "Things are moving to Texas," whereas the link with the preceding discourse is stronger if primary claim of sentence (3) is something like "Mr. and Mrs. Gallegos are moving to Texas."

The relation between sentences (2) and (3) is one of violated expectation. If someone enjoys some situation, as in sentence (2), then they will want to remain in that situation, and if someone wants something, then defeasibly they will get it. But sentence (3) describes a change of state out of that situation, violating our defeasible inference. In a violated expectation relation, the primary claim is usually the violation, in this case, the Gallegos couple's moving to Texas. Sentences (2) and (3) thus form a composite segment conveying the "moving to Texas" event. In the course of recognizing this relation, we resolve "they" to "Rosey Gallegos and her husband".

The relation between sentence (1) and sentences (2) and (3) is one of exemplification. In this relation an assertion is made or implied about a class of entities, and then the same assertion is made or implied about members of that class. The class is "those who want to stay". We recognize Rosey Gallegos and her husband as instances of that class, in part because it supports a coherence relation of exemplification, but more because sentence (2) tells us that they enjoy

living in Wyoming; enjoying something implies wanting it, and the "residing" sense of "living" is inferentially related to "stay". Struggling is having difficulty in achieving a goal. We learn the goal from what it is that is wanted, namely, the staying. Difficulties defeasibly cause failures to achieve one's goals. It is therefore a possible inference that some members of this class will not stay where they are, i.e., they will move.

Sentence (1) illustrates an interesting clause-internal coherence. There is a violated expectation relation between the relative clause "who want to stay" and the inference we have drawn from "those . . . are struggling." (Cf., Hobbs, 2008). This in fact parallels the structure of sentences (2) and (3).

Note that in discovering this coherence structure, we would have found a rich network of coreference. The struggling leads to the moving. The enjoying instantiates the wanting, the living here instantiates the staying, and the Gallegos couple instantiates the residents.

As the audio tells this story, the video tells a story that is related, but not closely related. In Alexandre et al. (2005), we developed a language (Video Event Representation Language, or VERL) for describing the structure of events in video data. The representation scheme is equivalent to first-order logic, and the language provides predicates, such as *subevent* and *change* that are useful for inferring large-scale events from observables. Several hundred event schemas have been encoded in VERL.
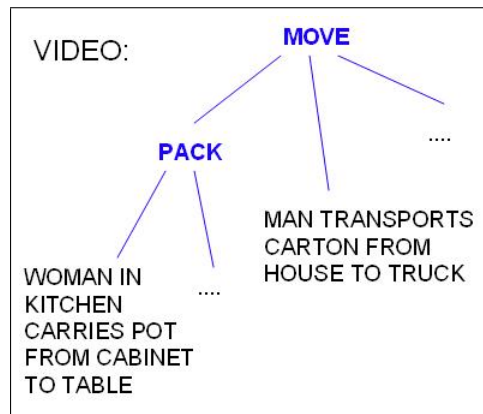
Fig. 3. VERL Representation of "Moving"

A VERL description of "moving", in the sense of changing residences, would have packing and loading a truck as subevents. Packing would involve changing the location of household objects to cartons. This is illustrated in Figure 3. Interpreting the video stream would amount to recognizing the woman moving the flowerpot from on top of the cupboard to the table as instantiating the

packing part of the schema, and recognizing the man wheeling the cartons from the house to the truck as instantiating the loading step in the schema.

But we are also looking for the single best interpretation for the discourse structure and the video structure together. This is achieved by identifying the moving schema with the second clause in sentence (3). The Gallegos couple is instantiating the moving scheme. In the course of making this identification, we identify the woman in the video as Rosey Gallegos. We take the kitchen we see to be the kitchen that is part of the house we see, and we take this house as referring to the same entity as is referred to by "here" in sentence (2) and the residence implicit in the word "residents" in sentence (1). We identify the man wheeling the cartons as Rosey Gallegos's husband (or perhaps an employee of the moving company).

4   Coreference

Determining object identity across frames of video data is a thriving area of research (e.g., Kang et al., 2003, 1005). This is essentially a coreference resolution problem. Two successive frames present images of the same kind of object, and we need to infer the identity or nonidentity of the subimages. This is often especially important when we are trying to detect changes in the camera angle and changes in the world. An object may be related to another object in one way in one frame and in a different way in the next frame. Recognizing this requires recognizing object identity across the two frames.
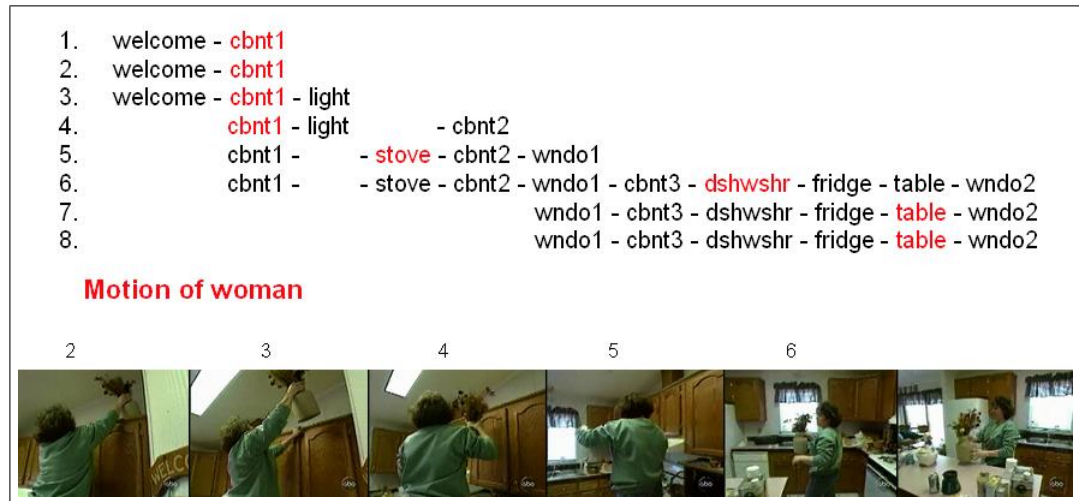

Fig. 4. Three seconds of the news broadcast video

Consider the stretch shown in Figure 4, of three seconds of the news broadcast video. The accompanying diagram labels the principal stationary objects visible in each frame, assuming we are able to label them as such and determine their identity across frames. In Frame 2 we see a "Welcome" sign and a cabinet. In Frame 3 we see these and the light in addition. In Frame 4 we no longer see the "Welcome" sign. In Frame 5 we no longer see the light, but the stove, another cabinet, and a window come into view. In Frame 6 a third cabinet, the dishwasher and fridge, and a table come into view. In Frame 7 the stove and two cabinets are gone, but a second window comes into view. If we were able to recognize the cross-frame coreference relations, we would be able to reconstruct the layout of the kitchen.

The woman is located at the first cabinet in Frames 2, 3, and 4. She is located at the stove in Frame 5, at the dishwasher in Frame 6, and at the table in Frame 7. A change in location is a "move", so recognizing the cross-frame identity of the woman and the location relations would enable us to infer that she had moved from the first cabinet to the table.
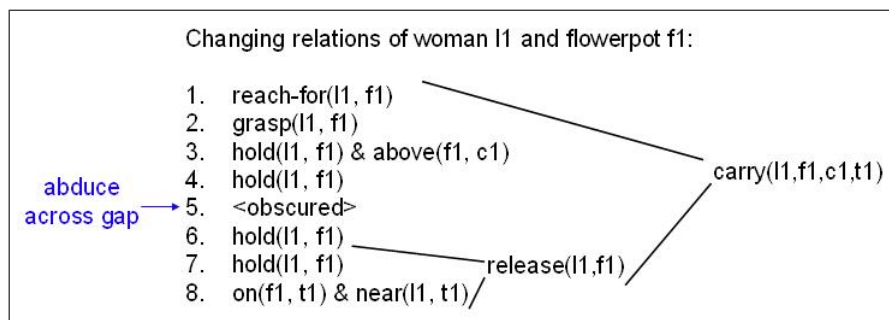

Fig. 5. Inference Illustrating a Woman Carrying a Vase

In Frame 1 (shown in Figure 1) she is reaching for the vase. In Frame 2 she is grasping it. In Frames 3, 4, and 6, she is holding it. We can't tell in Frame 5, but abductively we assume she is; we have no evidence to the contrary, and that tells the simplest story. In Frame 7, she is near the table, and the vase is nearly on the table. From Frame 6 to Frame 7 we (almost) have a change of state from holding the vase to not holding the vase, and this is a relese action. To grasp for something, hold it while moving, and then release it is to carry the object. Thus, by recognizing cross-frame identities along with changing relations between the objects, we would have been able to infer a carrying event by the woman of the vase from the first cupboard to the table. This inference is illustrated in Figure 5.

We would similarly be able to interpret the later part of the video stream as a man transporting a carton from the house to the truck.
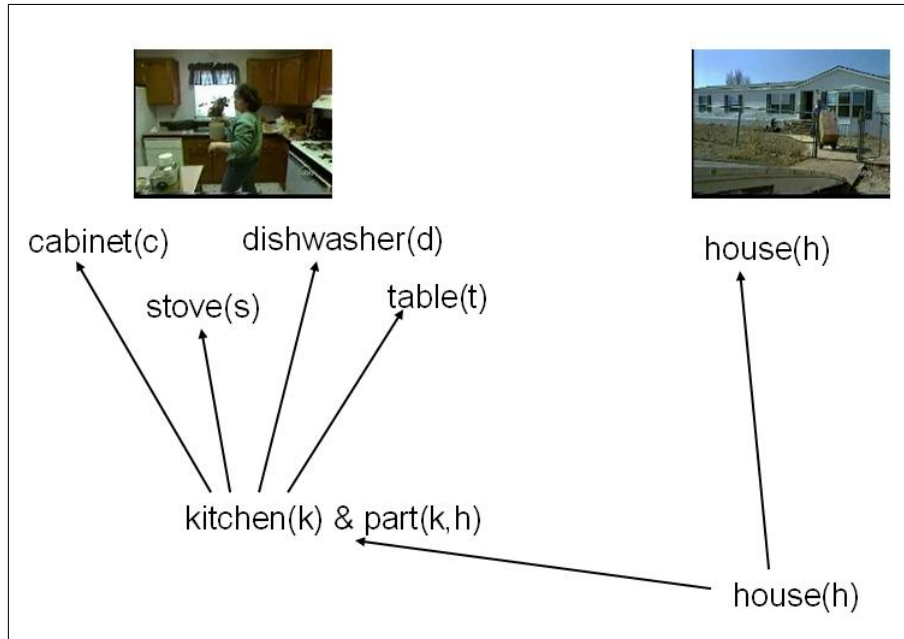
Fig. 6. Video-Video Coreference to Across Scenes

Figure 6 shows video-video coreference across scenes. We know that houses have kitchens, and that kitchens have cabinets, stoves, dishwashers, and tables. We know about the cabinets, stove, dishwasher, and table in the first scene, and about the house in the second scene. If we interpret the first scene as being in the kitchen that is part of the house shown in the second scene, we get the most economical interpretation of the video stream as a whole. Figure 7 is a screen shot of Mini-Tacitus solving this coreference problem, given the axioms indicated.

The examples so far have been of video-video coreference. A simple example of text-video coreference is seen in Frame 4. The video shows a woman holding a vase. The accompanying audio that half second is "Rosey Gallegos and ..." The text logical form contains $Rosey(r1) \wedge Gallegos(r1)$ and the video logical form contains $woman(w1)$. We know that Rosey is a woman's name:

$$(\forall x)[Rosey(x) \supset woman(x)]$$

We get the most economical explanation if we assume that Rosey is the woman being shown.

An example of text-text coreference arises in sentence (1). The word "struggling" has an implicit argument. When people struggle, there is something they are struggling for, but the sentence does not convey this with compositional semantics. The interpretation of this sentence is illustrated in Figure 8. The event

No.	Props	Backchained Props	Cost	Parent	Axiom Number
0	cabinet(c):0:30 , dishwasher(d):0:30 , stove(s):0:30 , table(t):0:30 , house(h):0:30 ,		150	0	
1	cabinet(c):1:0 , dishwasher(d):1:0 , stove(s):1:0 , table(t):1:0 , house(h):0:30 , kitchen(ax-x5):1:60 ,	cabinet(c):0:30 , table(t):0:30 , dishwasher(d):0:30 , stove(s):0:30 ,	90	0	ax1
2	**cabinet(c):2:0 , dishwasher(d):2:0 , stove(s):2:0 , table(t):2:0** , house(h):0:30 , **kitchen(ax-x5):2:0** ,	cabinet(c):1:0 , table(t):1:0 , dishwasher(d):1:0 , stove(s):1:0 ,	30	1	ax1
3	cabinet(c):1:0 , dishwasher(d):1:0 , stove(s):1:0 , table(t):1:0 , **kitchen(ax-x5):2:0 , part(ax-x5,ax-x2):2:6 , house(ax-x2):2:6** ,	kitchen(ax-x5):1:60 ,	12	1	ax2
4	**cabinet(c):2:0 , dishwasher(d):2:0 , stove(s):2:0 , table(t):2:0** , part(ax-x5,ax-x2):2:6 , house(ax-x2):2:6 , kitchen(ax-x5):2:0 ,	cabinet(c):1:0 , table(t):1:0 , dishwasher(d):1:0 , stove(s):1:0 ,	12	3	ax1

Lowest cost Interpretation

Fig. 7. Mini-TACITUS solving the Video-Video coreference Problem. "Props" stores all the video information currently in memory. Each Prop contains a "depth" (field after the first colon) and a "cost" (field after the second colon). The cost keeps decreasing as more Props can be explained by the Axioms. The lowest cost Interpretation is highlighted.
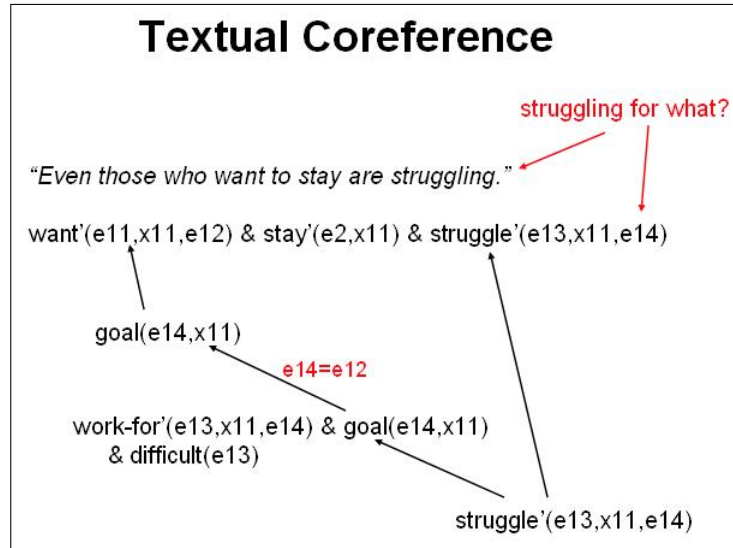
Fig. 8. Text-Text coreference across scenes

$e13$ is a struggling event by $x11$ for some unspecified state $e14$. If someone struggles for something, then that something is a goal that they work for, where the working for is difficult. If someone $x11$ has something $e14$ as a goal, then they want that goal. If we assume this wanting implicit in the struggle is identical to the wanting mentioned in the sentence, then we identify $e14$ with the object $e12$ of the wanting. We have thereby resolved the implicit argument of "struggling".

Figure 9 is a screenshot of MiniTacitus solving this problem and the problem of simultaneously identifying the woman in the video as Rosey Gallegos.

5 Interpreting Dynamically

We have been describing how the video and text are interpreted in concert after the fact, after all the information is in. But of course the way the viewers experience it is through time, and they construct their interpretations dynamically. In this section we describe how this might go, through 13 frames of the video, at half-second intervals, along with what is being spoken during that half second. For each step, we describe the video frame, the text, and then the interpretation that could be happening with the information so far available. (Nothing significant is happening for the first three frames.) This account is not implemented, but it is good to examine what kind of real-time interpretative processes we want to have eventually.

1. **Video:** Woman reaching for vase on top of cabinet.
   Text: Even those
2. **Video:** Woman grasping vase on top of cabinet.

No.	Props	Backchained Props	Cost	Parent	Axiom Number
0	want'(e1,x1,e2):0:30 , stay'(e3,x1):0:30 , struggle'(e4,x1,e5):0:30 ,		90	0	
1	**want'(e1,x1,e2):1:0** , stay'(e3,x1):0:30 , struggle'(e4,x1,e5):0:30 , **goal'(e2,x1):1:3** ,	want'(e1,x1,e2):0:30 ,	63	0	2
2	want'(e1,x1,e2):0:30 , stay'(e3,x1):0:30 , **struggle'(e4,x1,e5):1:0** , **work-for'(e4,x1,e5):1:3** , **goal'(e5,x1):1:3** , **difficult(e5):1:3** ,	struggle'(e4,x1,e5):0:30 ,	69	0	1
3	**want'(e1,x1,e2):1:0** , stay'(e3,x1):0:30 , **struggle'(e4,x1,e5):1:0** , **work-for'(e4,x1,e5):1:3** , **difficult(e5):1:3** , **goal'(e5,x1):1:3** ,	struggle'(e4,x1,e5):0:30 ,	39	1	1
4	**want'(e1,x1,e5):1:0** , stay'(e3,x1):0:30 , **struggle'(e4,x1,e5):1:0** , **work-for'(e4,x1,e5):1:3** , **difficult(e5):1:3** , **goal'(e5,x1):1:3** ,	want'(e1,x1,e5):0:30 ,	39	2	2

Coreferenced Arguments   Lowest Cost Interpretation

Fig. 9. Mini-TACITUS interpreting the sentence: "Those who want to stay are struggling." Interpretation number 4 is the lowest cost interpretation. The 3rd argument for "want" and "struggle" are now referring to the same "goal", which is the meaning conveyed in the sentence.

Text: who want to stay

3. Video: Woman lifting vase from top of cabinet.

    Text: are struggling.

4. Video: Woman next to cabinet holding vase.

    Text: Rosey Gallegos and

Here we can recognize that Rosey Gallegos is a woman and therefore possibly identical with the woman in the video. A woman is a person, and so are residents, so it is possible that Rosey Gallegos is an instance of the mentioned residents. That would make the exemplification relation a likely candidate for the coherence relation between the first sentence and the current sentence.

5. Video: Woman carries the vase past stove and sink.

    Text: her husband

The husband is another possible instance of the mentioned residents. From the stove and sink, we can infer that the woman is in a kitchen.

6. Video: Woman in kitchen carries vase past stove and toward table.

    Text: have enjoyed

The wanting of sentence (1) is linked with the enjoying of sentence (2).

7. Video: Woman in kitchen carries vase from cabinet to table.

    Text: living here for over

The living of this sentence is linked with the "stay" and "residents" of sentence (1). The word "here" is linked with the house which the kitchen can be presumed to be a part of.

8. Video: Woman places vase on table in kitchen.

    Text: a decade. But now

The word "But" signals a contrast or a violated expectation between sentences (2) and (3).

9. Video: Man wheeling cartons out of house.

    Text: his oil industry

We can identify the man in the video and the referent of "his" with Rosey Gallegos's husband. We infer that the house is the house the kitchen is a part of.

10. Video: Man wheeling cartons out of house.

    Text: job has been

No new inferences are possible at this point.

11. Video: Man wheeling cartons out of house.

    Text: moved to Texas

No new inferences are possible at this point.

12. Video: Man wheeling cartons out of house.

    Text: and they

The "and" tells us that one possible relation between the two clauses in sentence (3) is causality. We can tentatively resolve "they" to Rosey Gallegos and her husband since that is the most recent set referred to.

13. Video: Man wheels cartons up ramp onto truck.

    Text: are moving with it.

Heuristics (Hobbs, 1976) can resolve "it" to the husband's job. We verify the causal relation between the two clauses of sentence (3). This makes the Gallegos couple's moving the primary claim of sentence (3). We can then verify the violated expectation relation between sentence (2) and (3). Again the primary claim of the segment from (2) to (3) is the Gallegos couple's moving. This can then be seen as the unsuccessful outcome of the struggle to stay, in sentence (1), thereby verifying the exemplification relation between segment (1) and segment (2-3). At the same time, primed by "moving" in the second clause of sentence (3), the woman's carrying the vase from a high location to a more convenient lower location can be seen as an instance of packing, the man wheeling the cartons out can be seen as an instance of loading the truck, and the two together are subevents of the moving schema. We thereby infer a moving schema, and link it with the word "moving" in sentence (3).

6   Conclusion

The work described here is very preliminary. We have implemented a uniform representation for the content of video and text, and have developed reasonably reliable software for translating text into that notation. We have developed an abductive inference engine for interpreting the logical representations and linking them via coreference relations within and across modalities. We have not yet used it for recognizing large-scale events such as the Moving schema.

On the other hand, we have only tested this framework on a very small amount of data using a knowledge base of axioms that included only the knowledge needed for these examples. In a parallel effort we are building up a large knowledge base of commonsense knowledge geared to natural language understanding, and eventually we would like to see how well the framework scales up when this much knowledge is available. Finally, of course, work elsewhere continues to improve the speech and image recognition that this effort relies on.

Acknowledgments

References

1. Alexandre, Francois R.J., Ram Nevatia, Jerry R. Hobbs, and Robert C. Bolles, 2005. "VERL: An Ontology Framework for Representing and Annotating Video Events", IEEE Multimedia, Volume 12, October-December 2005, pp. 76-86.
2. Espinosa Peraldi, Sofia, Atila Kaya, Sylvia Melzer, Ralf Möller, Michael Wessel, 2007. "Towards a Media Interpretation Framework for the Semantic Web", in *Proceedings*, IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), No. 1331876, pp. 374-380. IEEE Computer Society, October 2007, Silicon Valley, USA.

3. Hobbs, Jerry R., 1978, "Resolving Pronoun References", *Lingua*, Vol. 44, pp. 311-338. Also in *Readings in Natural Language Processing*, B. Grosz, K. Sparck-Jones, and B. Webber, editors, pp. 339-352, Morgan Kaufmann Publishers, Los Altos, California.

4. Hobbs, Jerry R., 1985a. "On the Coherence and Structure of Discourse", Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

5. Hobbs, Jerry R. 1985b. "Ontological Promiscuity." *Proceedings, 23rd Annual Meeting of the Association for Computational Linguistics*, pp. 61-69. Chicago, Illinois, July 1985.

6. Hobbs, Jerry R. 2008. "Clause-Internal Coherence", to appear in P. Kuehnlein (Ed.), *Constraints in Discourse*.

7. Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. "Interpretation as Abduction", *Artificial Intelligence*, Vol. 63, Nos. 1-2, pp. 69-142.

8. Kang, Jinman, Isaac Cohen and Gérard Medioni, 2003. "Continuous Tracking Within and across Camera Streams," IEEE CVPR, 2003.

9. Kang, Jinman, Isaac Cohen, Gérard Medioni, 2005. "Persistent Objects Tracking Across Multiple Non Overlapping Cameras", IEEE WACV, pp. 112-119.

10. Shanahan, Murray P., 2005. "Perception as Abduction: Turning Sensor Data into Meaningful Representation", *Cognitive Science*, Vol. 29, pp. 103-134.