Supplementary Report on Caption Processing, Alignment, and Integration

1. Status and outline of the processing stages

We have reached the point where some captions are processed end-to-end, i.e., they are syntactically and semantically analyzed, simplified and canonicalized as much as possible, and used to generate inferences relevant to integration with image information, and this information is then used to align caption individuals with image individuals, and finally the information from the two sources in merged and presented in summary form.

In a little more detail, the stages of this completely automatic process can be described at a high level as follows (many details have been provided in previous reports). The references to "TTT" are to our new template-to-template (or tree-to-tree) transducer, which has been extremely helpful in enabling rapid, transparent implementation of several aspects of the linguistic and inferential processing.

1. Read a caption and corresponding image-derived data from files, converting the image-derived data to a Lispified form suitable for integration with caption-derived information;
2. Parse the caption, and refine and repair the parse tree (using TTT and Lisp code) to better enable interpretation;
3. Derive an unscoped logical form (ULF) from the parse tree using compositional interpretive rules;
4. Resolve simple anaphors in the ULF (e.g., replacing a reference to "her" with a reference to Tanya);
5. Resolve several types of ambiguity in logical structure, including quantifier scope, and/or scope, and tense meaning (the last is not relevant to the captions processed so far);
6. Reduce temporary LF keywords, and repair the LF repair and simplify it into a canonical form suitable for EPILOG input; this is mostly done with TTT rules;
7. Store the canonical formulas in EPILOG;
8. Identify caption entities, and then likely persons among those entities, along with properties inferrable from the names (e.g., "Grandma Lillian" indicates that this is a grandmother); this is based on name gazetteers and knowledge about titles;

9. Store the name-derived facts in EPILOG, and ask EPILOG to assign certainties to several dozen propositions about each caption person, such as their gender, age, hair color, and other properties relevant to correlating the caption information with the image-derived information.

10. Use the information thus obtained to align caption persons with image persons; the alignment algorithm calculates a mismatch score between caption individuals and image individuals, based in part on confusion probabilities between related alternatives such as being male or female, brown-haired or blond (among other alternatives), being in one or another age group, etc. The mismatch score also depends on the certainty with which the image-derived and caption-derived properties are inferred. (There are technical reasons having to do with the variability in the number of properties inferrable for caption individuals that make mismatch scores more suitable than similarity scores for finding plausible alignments.)

11. Use the best alignment that has been found to merge all the known properties of the individuals found. At this point we can also begin to ask EPILOG relational questions such as "Is some individual Reggi's friend?", or "Who is in the park?".

2. Two examples

Two examples that have been fully processed are those shown in figues 1 and 2.

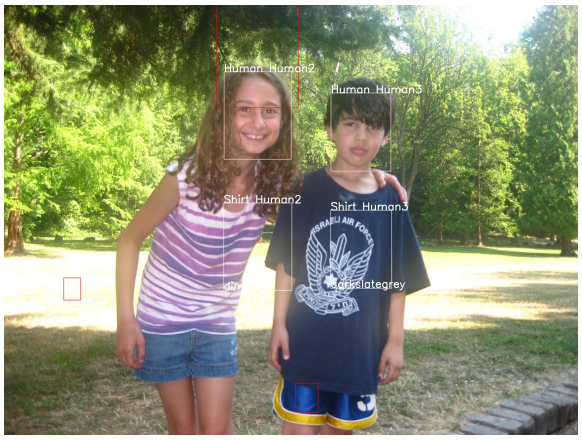

Fig. 1 *"Ben and his friend Reggi, at the park."*

Fig. 2 *"Grandma Lillian and Uncle Moshe at Tanya's graduation party."*

The first example is relatively simple, since if the gender of Reggi (Human2 in the image) and Ben (Human3 in the image) can be correctly identified in both the image and the caption, then there can be little doubt about the correct alignment. The second example is more challenging, because one of the individuals mentioned in the caption, namely Tanya, does not appear in the image, and furthermore, it turns out that the image-derived information about Uncle Moshe (or rather, Human4) assigns only

slightly higher probability to male gender than to female gender, and is otherwise consistent with a female (e.g. absence of facial hair).

2.1 Caption processing for Figure 1

In some detail, then, here is how the processing proceeded for the two images, starting with "Reggi-Ben.jpg".

Step 1 of the above outline produced the following Lispified image information:

(REGGI-BEN.JPG

(HUMAN2 PERSON.N 0.99 BABY-OR-TODDLER.N 0.0833 CHILD.N 0.2403 TEEN.N 0.2415 YOUNG-ADULT.N 0.0803 YOUNGISH-ADULT.N 0.0411 MIDDLE-AGED.A 0.1553 SENIOR.N 0.1582 RACIALLY-WHITE.A 0.3312 RACIALLY-BLACK.A 0.1508 RACIALLY-ASIAN.A 0.2048 RACIALLY-HISPANIC.A 0.1071 RACIALLY-EAST-INDIAN.A 0.206 RACIALLY-ARABIC.A 0 MALE.N 0.215 FEMALE.N 0.785 BALD-HEADED.A 0.0832 BLOND-HAIRED.A 0.2397 BROWN-HAIRED.A 0.2683 DARK-HAIRED.A 0.1686 GRAY-HAIRED.A 0.1132 RED-HAIRED.A 0.127 BLUE-EYED.A 0.2625 BROWN-EYED.A 0.2143 DARK-EYED.A 0.3147 GREEN-EYED.A 0.2085 THIN-FACED.A 0.0735 OVAL-FACED.A 0.5641 ROUND-FACED.A 0.3624 MUSTACHELESS.A 1 MUSTACHIOED.A 0 BEARDLESS.A 1 BEARDED.A 0 NOT-WEARING-GLASSES.A 0.6429 WEARING-GLASSES.A 0.3571 NOT-WEARING-SUNGLASSES.A 1 WEARING-SUNGLASSES.A 0 UNSMILING.A 0 SMILING.A 1 HATLESS.A 1 WEARING-A-HAT.A 0)

(HUMAN3 PERSON.N 0.99 BABY-OR-TODDLER.N 0.1555 CHILD.N 0.2151 TEEN.N 0.1911 YOUNG-ADULT.N 0.1141 YOUNGISH-ADULT.N 0.0711 MIDDLE-AGED.A 0.1403 SENIOR.N 0.1128 RACIALLY-WHITE.A 0.3423 RACIALLY-BLACK.A 0.0537 RACIALLY-ASIAN.A 0.1841 RACIALLY-HISPANIC.A 0.1955 RACIALLY-EAST-INDIAN.A 0.2244 RACIALLY-ARABIC.A 0 MALE.N 0.6073 FEMALE.N 0.3927 BALD-HEADED.A 0.1124 BLOND-HAIRED.A 0.1925 BROWN-HAIRED.A 0.2171 DARK-HAIRED.A 0.2305 GRAY-HAIRED.A 0.1552 RED-HAIRED.A 0.0923 BLUE-EYED.A 0.3275 BROWN-EYED.A 0.1777 DARK-EYED.A 0.3203 GREEN-EYED.A 0.1745 THIN-FACED.A 0.2956 OVAL-FACED.A 0.5548 ROUND-FACED.A 0.1496 MUSTACHELESS.A 1 MUSTACHIOED.A 0 BEARDLESS.A 0.9047 BEARDED.A 0.0953 NOT-WEARING-GLASSES.A 0.8647 WEARING-GLASSES.A 0.1353 NOT-WEARING-SUNGLASSES.A 1 WEARING-SUNGLASSES.A 0 UNSMILING.A 0.6186 SMILING.A 0.3814 HATLESS.A 0.5016 WEARING-A-HAT.A 0.4984))

Note the separate lists of properties, with probabilities, of Human2 (Reggi) and Human3 (Ben), whose identities of course are not yet known. Note also that we have, in effect, probability distributions over "spectra" of properties (a term borrowed from Rudolf Carnap's work on inductive probability), such as binary spectra like gender, and multi-element spectra like age group and hair color. Parsing of the caption, using the Charniak parser and some slight refinement using TTT rules (step 2 above) yields the parse tree

```
(S1
  (NP
    (NP (NP (NNP BEN)) (CC AND)
          (NP (DETP (NP (PRP HE)) (POS |'S|)) (NN FRIEND) (NNP REGGI)))
  (|,| |,|) (PP-AT (IN AT) (NP (DT THE) (NN PARK))) (\. \.)))
```

So we have an NP (noun phrase) consisting of two constituent NPs (one for each person) coordinated with "and", and followed by a PP-AT (prepositional phrase using preposition "at"). As a "sentence", this is flawed, as it lacks a verb. This is dealt with later in logical-form processing.

The initial, unscoped logical form (ULF) computed in phrase-by-phrase compositional fashion from this parse tree, using around a hundred interpretive rules (each of which can handle multiple variants of certain phrase types) is the following (as per step 3 above):

```
(:A
 (:F SET-OF BEN.NAME
  (:Q THE.DET
   (:L  X (:I  (:I X = (:Q THE.DET (:L X (:I (:I X FRIEND.N) AND (:I X PERTAIN-TO HE.PRO)))))
           AND (:I X = REGGI.NAME)))))
 (:P AT.P (:Q THE.DET PARK.N)))
```

The following paragraph can be skipped by readers not concerned with details of the logical form syntax.

The colon-prefixed letters in this ULF are keywords that indicate the types of logical constituents they head. In particular, :A indicates an augmentation structure (here the augmentation of the description of a set of two individuals with a locative predicate, where the latter expresses the property of being at the park). The :F indicates a function application (namely application of SET-OF to the individuals described as BEN.NAME and "the friend pertaining to him" (with the anaphor "HE.PRO" still unresolved). The :Q indicates 3 occurrences of unscoped quantifier THE.DET. (An unscoped quantifier simply pairs a quantifier with a predicate, as in (:Q THE.DET PARK.N), not yet introducing a quantified variable). The :L is the lambda abstraction operator (which here forms the property of "being identical with the friend that pertains to HE.PRO and is identical with REGGI.NAME"). The :I is used to indicate an infix sentential formula, i.e., one where the subject argument precedes the predicate, e.g., as in (:I X PERTAIN-TO HE.PRO). Finally, the :P indicates predicate application to an argument, in this case, application of the AT.P relational predicate to the (unscoped) argument (:Q THE.DET PARK.N).

The resolution of anaphors (step 4) using TTT rules slightly alters the obove ULF, replacing HE.PRO with BEN.NAME based on their positions, gender agreement and number agreement (singular). The following shows the result, after an additional step (part of step 5) that incorporates the augmentation structure (signaled by keyword :A) as a predication -- hence the initial infix keyword :I.

```
(:I
 (:F SET-OF BEN.NAME
  (:Q THE.DET
   (:L X  (:I (:I X = (:Q THE.DET  (:L X (:I (:I X FRIEND.N) AND (:I X PERTAIN-TO BEN.NAME)))))
              AND (:I X = REGGI.NAME)))))
 (:P AT.P (:Q THE.DET PARK.N)))
```

The rest of step 5 scopes the THE.DET quantifiers so that they now are positioned outside the formulas that introduced them (as unscoped arguments of predicates and equality):

```
(THE.DET Y
 (THE.DET Z (:I (:I Z FRIEND.N) AND (:I Z PERTAIN-TO BEN.NAME))
  (:I (:I Y = Z) AND (:I Y = REGGI.NAME)))
 (THE.DET V (:I V PARK.N)
  (:I (:F SET-OF BEN.NAME Y) (:L W (:I W AT.P V)))))
```

Each quantifier now also binds a variable, and the variable occupies the argument position where the quantifier was introduced. This is an ambiguity resolution step, because decisions need to be made about the relative scopes of the quantifiers. This is done by heuristic rules that are an integral part of the scoping algorithm. The scoping algorithm also assigns sentential scopes to unscoped occurrences of "and", "or", "but", and some other coordinators, but in this example there are no such occurrences (the "and" joining the two NPs was interpreted as a set-forming function, rather than as sentential conjunction).

We skip the part of step 5 that would interpret tense in a tensed sentence, and associates a communicative act with the caption (i.e., information is being communicated to a recipient), because the caption is untensed, and the communicative act is discarded again (by TTT rules) as unnecessary for our immediate purposes.

The final transformations of the LF in step 6
- eliminate unnecessary keywords (all but :L, the lambda abstractor),
- interpret "is a friend that pertains to" as "is a friend-of",
- Skolemize existential and definite quantifiers (i.e., introduce new names for the entities that are presumed to exist; the new names are suffixed with ".SK"),
- eliminate top-level conjunctions, and
- substitute proper names for Skolem entities wherever this is made possible by an equation relating a proper name to a Skolem entity.

These miscellaneous transformations are all implemented using simple TTT rules. The result is

```
((REGGI.NAME FRIEND-OF.N BEN.NAME) (PARK1.SK PARK.N)
 ((SET-OF BEN.NAME REGGI.NAME) AT.P PARK1.SK)).
```

i.e., we have a list of formulas stating that Reggi is a friend of Ben, some entity with Skolem name PARK1.SK is a park, and the set of Ben and Reggi are at the park.

These formulas are now stored in EPILOG, and this incidentally immediately allows some inferential question answering. For example, the question

(BEN.NAME FRIEND-OF.N REGGI.NAME)

("Is Ben a friend of Reggi?") is answered affirmatively, because the EPILOG KB contains an axiom stating the friendship is symmetric (if x is a friend-of y then y is a friend-of x).

The next step (8) is to identify caption entities

(BEN.NAME REGGI.NAME PARK1.SK)

(Ben, Reggi, and the park), to select the likely persons

(BEN.NAME REGGI.NAME)

from this list, and to infer properties of these persons indicated by their names. As might be expected, these properties are

(BEN.NAME MALE.N), (BEN.NAME PERSON.N), and
(REGGI.NAME FEMALE.N), (REGGI.NAME PERSON.N).

and these are asserted into EPILOG (the propositions are hedged slightly, using probability .99).

The next step (9) brings to bear EPILOG's KB and inference capabilities. The KB is still small at this point (137 axioms), but contains enough knowledge about family relationships, gender, age, hair color, facial hair, and glasses to be able to make inferences relevant to aligning caption persons with image persons. It also contains a few miscellaneous items about graduation parties, camp, and clothing. For the sparse caption information under consideration, EPILOG draws only two new conclusions (apart from the gender information it already has about Ben and Reggi), namely

(REGGI.NAME MUSTACHELESS.A), (REGGI.NAME BEARDLESS.A),

which are inferences from Reggi's female gender. EPILOG cannot draw this conclusion about Ben, because the caption provides no clue that Ben is a child.

Then comes the crucial step (10) of aligning the caption individuals Reggi and Ben with the image individuals Human2 and Human3. The above inferences about Reggi's lack of facial hair neither help nor hinder the alignment; the gender information alone leads to the correct alignment, since fortunately the image processing provided fairly firm gender information about Human2 (MALE.N 0.215, FEMALE.N 0.785) and Human3 (MALE.N 0.6073, FEMALE.N 0.3927):

((BEN.NAME HUMAN3 -0.5) (REGGI.NAME HUMAN2 -0.5)).

(The discrepancy scores of -.5 need not concern us.) Finally (step 11) the caption information is merged with the image-derived information. In this process, properties that receive both image-based probabilities and caption-based probabilities have their probabilities combined (accordng to a disjunctive rule), and in addition, low-

probability image-derived properties are weeded out, and the most likely one or two alternatives in each spectrum of properties have their probability boosted (where the amount of boosting depends on the distribution of probabilities over each spectrum). The result of merging for the Reggi-Ben.jpg image and caption is presented in this summary form:

(REGGI-BEN.JPG

((REGGI.NAME HUMAN2 0.6666667) (PERSON.N 0.99995) (FEMALE.N 0.998925)
(MUSTACHELESS.A 1.0) (BEARDLESS.A 1.0) (TEEN.N 0.37075)
(CHILD.N 0.37015) (RACIALLY-WHITE.A 0.6656) (BROWN-HAIRED.A 0.38415)
(BLOND-HAIRED.A 0.36985) (DARK-EYED.A 0.40735) (BLUE-EYED.A 0.38125)
(OVAL-FACED.A 0.78205) (NOT-WEARING-GLASSES.A 0.82145)
(NOT-WEARING-SUNGLASSES.A 1.0) (SMILING.A 1.0) (HATLESS.A 1.0))

((BEN.NAME HUMAN3 0.6666667) (PERSON.N 0.99995) (MALE.N 0.9980365)
(CHILD.N 0.35755) (TEEN.N 0.34555) (RACIALLY-WHITE.A 0.67114997)
(DARK-HAIRED.A 0.36525) (BROWN-HAIRED.A 0.35855)
(BLUE-EYED.A 0.41375) (DARK-EYED.A 0.41015) (OVAL-FACED.A 0.7774)
(MUSTACHELESS.A 1.0) (BEARDLESS.A 0.95235)
(NOT-WEARING-GLASSES.A 0.93235004) (NOT-WEARING-SUNGLASSES.A 1.0)
(UNSMILING.A 0.8093) (HATLESS.A 0.5008) (WEARING-A-HAT.A 0.4992)))

Note that these descriptions are still ambivalent about whether Reggi and Ben are children or teens, and also about hair color (brown or blond in Reggi's case, brown or dark in Ben's case) and eye color (blue-eyed or dark-eyed). This is reasonable given the original image-based probability distributions.

2.2 Caption processing for Figure 2

Here we will be less verbose, having explained the processing steps for the first image in considerable detail.

For the Grandma-Moshe.jpg image the Lispified image-derived properties are as follows:

(GRANDMA-MOSHE.JPG

(HUMAN4 PERSON.N 0.99 BABY-OR-TODDLER.N 0.095 CHILD.N 0.1784 TEEN.N
0.2082 YOUNG-ADULT.N 0.0854 YOUNGISH-ADULT.N 0.0206 MIDDLE-AGED.A
0.2254 SENIOR.N 0.187 RACIALLY-WHITE.A 0.3844 RACIALLY-BLACK.A 0.0802
RACIALLY-ASIAN.A 0.1757 RACIALLY-HISPANIC.A 0.144
RACIALLY-EAST-INDIAN.A 0.2156 RACIALLY-ARABIC.A 0 MALE.N 0.5312
FEMALE.N 0.4688 BALD-HEADED.A 0.0428 BLOND-HAIRED.A 0.148
BROWN-HAIRED.A 0.2923 DARK-HAIRED.A 0.3059 GRAY-HAIRED.A 0.1021
RED-HAIRED.A 0.109 BLUE-EYED.A 0.2736 BROWN-EYED.A 0.3247 DARK-EYED.A
0.2677 GREEN-EYED.A 0.134 THIN-FACED.A 0 OVAL-FACED.A 0.5644
ROUND-FACED.A 0.4356 MUSTACHELESS.A 0.9049 MUSTACHIOED.A 0.0951
BEARDLESS.A 1 BEARDED.A 0 NOT-WEARING-GLASSES.A 0.9012

WEARING-GLASSES.A 0.0988 NOT-WEARING-SUNGLASSES.A 1
WEARING-SUNGLASSES.A 0 UNSMILING.A 0.1387 SMILING.A 0.8613 HATLESS.A
0.9373 WEARING-A-HAT.A 0.0627)

(HUMAN5 PERSON.N 0.99 BABY-OR-TODDLER.N 0.075 CHILD.N 0.1764 TEEN.N
0.1844 YOUNG-ADULT.N 0.0725 YOUNGISH-ADULT.N 0.0115 MIDDLE-AGED.A
0.2187 SENIOR.N 0.2616 RACIALLY-WHITE.A 0.3632 RACIALLY-BLACK.A
0.1346 RACIALLY-ASIAN.A 0.1661 RACIALLY-HISPANIC.A 0.1389
RACIALLY-EAST-INDIAN.A 0.1972 RACIALLY-ARABIC.A 0 MALE.N 0.3656
FEMALE.N 0.6344 BALD-HEADED.A 0.1539 BLOND-HAIRED.A 0.2138
BROWN-HAIRED.A 0.1387 DARK-HAIRED.A 0.1398 GRAY-HAIRED.A 0.3162
RED-HAIRED.A 0.0376 BLUE-EYED.A 0.2944 BROWN-EYED.A 0.206 DARK-EYED.A
0.2943 GREEN-EYED.A 0.2053 THIN-FACED.A 0 OVAL-FACED.A 0.4474
ROUND-FACED.A 0.5526 MUSTACHELESS.A 0.9049 MUSTACHIOED.A 0.0951
BEARDLESS.A 0.9047 BEARDED.A 0.0953 NOT-WEARING-GLASSES.A 0.0571
WEARING-GLASSES.A 0.9429 NOT-WEARING-SUNGLASSES.A 0.7782
WEARING-SUNGLASSES.A 0.2218 UNSMILING.A 0 SMILING.A 1 HATLESS.A
0.8107 WEARING-A-HAT.A 0.1893))

Note the very uncertain gender judgement (MALE.N 0.5312, FEMALE.N 0.4688) for Human4 (who is actually Uncle Moshe). This creates a risk that Human4 will be aligned with Tanya (who is also mentioned in the caption), but the slightly higher probability for MALE.N averts this error.

The parse tree after step 2 is

```
(S1
 (NP (NP (NP (NNP GRANDMA LILLIAN)) (CC AND) (NP (NNP UNCLE MOSHE)))
  (PP-AT (IN AT)
   (NP (DETP (NP (NNP TANYA)) (POS |'S|)) (NN GRADUATION) (NN PARTY)))
 (\. \.)))
```

This actually involved application of a repair rule, because the original Charniak parse failed to create separate NPs for Grandma Lillian and Uncle Moshe -- which would have led to interpretive errors.

The logical form after steps 3-5, and the keyword-deletion step of step 6, is

```
(THE.DET Y
  ((Y ((NN GRADUATION.N) PARTY.N)) AND (Y PERTAIN-TO TANYA.NAME))
  ((SET-OF GRANDMA_LILLIAN.NAME UNCLE_MOSHE.NAME)
   (:L Z (Z AT.P Y))))
```

Completion of step 6 then yields the canonical list of formulas (which are then stored in EPILOG in step 7)

```
((PARTY0.SK ((NN GRADUATION.N) PARTY.N))
 (PARTY0.SK PERTAIN-TO TANYA.NAME)
 ((SET-OF GRANDMA_LILLIAN.NAME UNCLE_MOSHE.NAME) AT.P PARTY0.SK)).
```

So this states that a certain entity with Skolem name PARTY0.SK is a graduation party that pertains to Tanya, and Grandma Lillian and Uncle Moshe (as a set) are at that party.

The analysis of the two names (step 8) produces the (somewhat redundant) information that

(GRANDMA_LILLIAN.NAME GRANDMOTHER.N),
(GRANDMA_LILLIAN.NAME FEMALE.N), (GRANDMA_LILLIAN.NAME PERSON.N)

and

(UNCLE_MOSHE.NAME UNCLE.N),
(UNCLE_MOSHE.NAME MALE.N), (UNCLE_MOSHE.NAME PERSON.N).

Upon storage in EPILOG, the following new information is inferred by EPILOG (step 9):

(TANYA.NAME TEEN.N) or (TANYA.NAME YOUNG-ADULT.N),
(TANYA.NAME MUSTACHELESS.A), (TANYA.NAME BEARDLESS.A)
(GRANDMA_LILLIAN.NAME SENIOR.N)
(or possibly middle-aged, with lower probability, or
possibly a youngish adult, with still lower probability)
(GRANDMA_LILLIAN.NAME GRAY-HAIRED.A)
(or brown-haired, with lower probability)
(GRANDMA_LILLIAN.NAME MUSTACHELESS.A)
(GRANDMA_LILLIAN.NAME BEARDLESS.A)
(GRANDMA_LILLIAN.NAME WEARING-GLASSES.A) (with .5 probability)

Based on all the caption-derived and inferred information, alignment with the image-derived information (step 10) yields

((GRANDMA_LILLIAN.NAME HUMAN5 -0.5) (UNCLE_MOSHE.NAME HUMAN4 -0.5)),

which is the correct alignment. The inferred knowledge that Grandma Lillian is probably gray-haired helped to ensure that Grandma Lillian, rather than Tanya, would be identified with Human5.

The final merger of information (step 11) yields the summary

(GRANDMA-MOSHE.JPG

((UNCLE_MOSHE.NAME HUMAN4 0.6666667) (PERSON.N 0.99995)
(MALE.N 0.997656) (MIDDLE-AGED.A 0.3627) (TEEN.N 0.3541)
(RACIALLY-WHITE.A 0.6922) (DARK-HAIRED.A 0.40295)
(BROWN-HAIRED.A 0.39615) (BROWN-EYED.A 0.41235) (BLUE-EYED.A 0.3868)
(OVAL-FACED.A 0.7822) (MUSTACHELESS.A 0.95245004) (BEARDLESS.A 1.0)
(NOT-WEARING-GLASSES.A 0.9506) (NOT-WEARING-SUNGLASSES.A 1.0)
(SMILING.A 0.93065) (HATLESS.A 0.96865))

((GRANDMA_LILLIAN.NAME HUMAN5 0.6666667) (PERSON.N 0.99995)
(FEMALE.N 0.99817204) (SENIOR.N 0.96308) (GRAY-HAIRED.A 0.924782)
(MUSTACHELESS.A 0.99881124) (BEARDLESS.A 0.9995235)
(WEARING-GLASSES.A 0.98593915) (RACIALLY-WHITE.A 0.6816)
(BLUE-EYED.A 0.3972) (DARK-EYED.A 0.39714998) (ROUND-FACED.A 0.7763)
(NOT-WEARING-SUNGLASSES.A 0.88909996) (SMILING.A 1.0)
(HATLESS.A 0.90534997)))

This again leaves some spectra ambivalent, but benignly so: Uncle Moshe has some probability of being middleaged or being a teen, which might lead to guessing an in-between age (20s or 30s). As in the first image, the hair color and eye color for the younger individual remain somewhat ambivalent. The "hatless" property for Uncle Moshe is an error, but an understandable one, caused by the difficulty of detecting caps in frontal views.

3. Comments on the significance of the results

These results are beginning to demonstrate, for the first time, the possibility of full comprehension of simple picture captions in the "family album" domain (rather than mere statistical associations between bag-of-words or phrasal features of captions with image features), and the use of such comprehension to align caption-derived entities with image-derived entities.

This is important, because it obviates the need for training of machine learning algorithms on massive picture+caption corpora annotated with alignment information -- which incidentally would be hard to obtain because very little in the way of family photos with captions (giving names of individuals) exists in the public domain; privacy issues make accumulation of such data very difficult. Besides this, even if such annotated data became available, application of current statistical and ML techniques would not lead to the kinds of inferencing capabilities that are a feature of our EPILOG-based and TTT-based symbolic approach. We noted the ability to infer relationships, and we are only some fairly small steps away from being able to answer questions, posed in English, such as "Does Ben have any friends?", "How old is Tanya?" (note: she appears in neither image), "Who was at the graduation party?", "Whose grandmother is Grandma Lillian?", or "Who was the graduate at the graduation party?" None of this information is directly supplied by the captions, let alone the images.

4. The next steps

We are still in the early stages of this work, and the immediate goal is of course to process a significant number of images. This will certainly require significant extensions of the EPILOG knowledge base to cover more properties of and relationships among concepts of ordinary language, particularly those often referenced in image captions (various types of people, pets, activities, special events, indoor and outdoor settings, etc.). However, we should emphasize that the techniques we have applied to the examples tested so far are very general: The parsing is completely

general; the compositional interpretive approach, though still unreliable in dealing with some phrase types (such as clausal adverbials) is also completely domain-independent; the name knowledge we have coded relies on quite large gazetteers and general heuristic routines; and the inference techniques used by EPILOG are also completely domain independent (and comprehensive, except that forward inference, unlike backward inference, is not yet adequate). Our new TTT tool is also perfectly general, and will facilitate various additional ways of refining, disambiguating and repairing parse trees and semantic formulas in future, as well as providing new methods for forward inference, question interpretation, and English generation.

In light of this, our infrastructure work will focus on further knowledge accumulation, forward inference mechansims, development and large-scale testing of TTT syntactic repair rules on large general English corpora. We are also working towards systematic interfaces between the image processing and caption processing components, and user interfaces that will enable easy uploading of images and captions, and prompting the system for the desired kinds of integration to be performed, and for the desired kinds of information to be returned.