Describing objects in visual scenes
Is visual salience like conversational salience?

Micha Elsner
Hannah Rohde, Alasdair Clarke

Department of Linguistics
The Ohio State University
University of Edinburgh
“Describe the person in the box so that someone could find them”
To the right of the men smoking a woman wearing a yellow top and red skirt.

woman in yellow shirt, red skirt in the queue leaving the building

the woman in a yellow short just behind the spray of the hose

Between the yellow and white airplanes there is a red vehicle spraying people with a hose. The people getting sprayed have a small line behind them. In the line there is a woman with brownish red hair, a yellow shirt and a red skirt holding a purse. She is standing behind a man dressed in green.
“The woman standing near the jetway”

- Overall target:
  - “the woman”
- Landmark:
  - “the jetway”
  - relative to “woman”
Motivation:

- Information structure via *discourse salience*:
  - Familiar / important / in common ground
- Image understanding via *visual salience*:
  - Perceptually apparent / attracts attention

What do they have in common?

This study:

- Complex information structure of relational descriptions
- Visual features matter...
- Visual salience *is* like discourse salience
Overview

Ordering strategies in the corpus

“Where’s Wally”: the dataset

Learning to use visual features

Experiments: predicting the order
Ordering strategies: direction

The woman standing near the jetway

Near the hut that is burning, there is a man...

Man... next to railroad tracks wearing a white coat

- Orders defined WRT first mention
- Information structure, not syntax
Basic ordering

- **RIGHT** default for landmarks (40%)
- **LEFT** default for image regions (57%)
  - “On the left is a woman”...
- Other orders are marked:
  - **LEFT** landmarks (33%)
  - **INTER** landmarks (27%)
Non-relational mentions

Look at the plane. This man is holding a box that he is putting on the plane.

- First mention isn’t relational
  - “There is”, “look at”, “find the”...
- Annotated as ESTABLISH construction
- Usually occurs with LEFT ordering
Where’s Wally: the dataset

By Martin Handford: Walker Books, London

- Published in US as “Where’s Waldo”
- Series of childrens’ books: a game based on visual search
- Gathered referring expressions through Mechanical Turk
- Each subject saw a single target in each image
28 images x 16 targets x 10 subjects per image
Why Wally?

- Wide range of objects with varied visual salience
- Deliberately difficult visual search
- Relational descriptions a must
  - Not: “Wally is wearing a red striped shirt and a bobble hat”
- Previous studies used fewer objects
- Got fewer relational descriptions

(Viethen+Dale ‘08)
The <targ>man</targ> just to the left of the <lmark rel="targ" obj="(id)">burning hut</lmark> <targ>holding a torch and a sword</targ>
Individual variation

For head/landmark pairs mentioned by multiple subjects:

- 65% agreement about mention direction
- 40% ESTABLISH constructions agreed on

Strategies are predictable but vary

- Based on other landmarks selected?
- Different cognitive strategies?
Visual information:

- Root area of object...
- (Low-level) visual salience of object
- Distance between objects

Visual salience:

- Psychological models of low-level vision
  (Toet ‘11, Itti+Koch ‘00, others)
- Where will people look in an image?
- Which objects are easy to find?
Salience map

- Based on responses from filter bank
- Bottom-up part of (Torralba+al ‘06)
Modeling: tag induction

- Information structure as tagging problem
- Each object has (hidden) type
  - Analogous to part of speech
- Order controlled by types

The woman standing near the jetway

![Diagram]

- target1
- landmark2
- right
- The woman standing near the jetway
Begin with simple discriminative system

- Features: discretized area, salience, distance
  - Thresholds set at training set quartiles
- Number of landmarks used for each object

The woman standing near the jetway
Multilayer system

- No longer reliant on hand-tuned discretization
- CRF/Neural Net with latent *type* variables
- Area, salience, deps predict type
- ...which predict direction

*The woman* standing near *the jetway*
Tag induction: *almost* grammar induction
   ▶ Not hierarchical yet though
Based on Berkeley-style latent variable grammar
   ▶ (Matsuzaki+al ‘05, Petrov+al ‘06,’08)
Implemented with Theano package
   ▶ Automatic computation of gradients
Visualization of types for objects
Linguistic analysis

- **Red**: smallest and hardest to see
  - Right > inter > left
- **Blue**: small
  - Right > inter > left
  - A few ESTABLISH
- **Green**: midsized
  - Left > inter = right
  - Common as ESTABLISH
- **Purple**: largest
  - Inter > left = right
Information ordered by givenness/familiarity:
(Prince ‘81, Birner+Ward ‘98 etc)

- Subject position: more familiar entities
- New information (outside common ground) later in sentence

Obama (given) has a dog named Bo (new)

- ESTABLISH construction introduces hearer-new entity (Ward+Birner ‘95)

Hey, look! There’s a huge raccoon asleep under my car (new)! (WB95 ex. 9)
Visual salience is similar:

- Highly visible landmarks appear left/inter
  - Treated as familiar entities
  - Assumed in common ground
- Harder-to-see landmarks on right
  - Assumed discourse-new
- ESTABLISH construction used for mid-sized entities
  - Used to place them on the left
  - Might not normally be on the left (not in common ground)
  - But are visually salient enough to motivate leftward order
Predicting the order

- Input: unordered abstract structure

<table>
<thead>
<tr>
<th>Acc (direction)</th>
<th>F (ESTABLISH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All RIGHT</td>
<td>36</td>
</tr>
<tr>
<td>Regs LEFT</td>
<td>43</td>
</tr>
</tbody>
</table>
Predicting the order

- Input: unordered abstract structure

<table>
<thead>
<tr>
<th></th>
<th>Acc (direction)</th>
<th>F (ESTABLISH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All RIGHT</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>Regs LEFT</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>Basic discr</td>
<td>50</td>
<td>43</td>
</tr>
<tr>
<td>Multilevel</td>
<td>52</td>
<td>50</td>
</tr>
</tbody>
</table>
Predicting the order

- Input: unordered abstract structure

<table>
<thead>
<tr>
<th></th>
<th>Acc (direction)</th>
<th>F (ESTABLISH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All RIGHT</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>Regs LEFT</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>Basic discr</td>
<td>50</td>
<td>43</td>
</tr>
<tr>
<td>Multilevel</td>
<td>52</td>
<td>50</td>
</tr>
<tr>
<td>Majority oracle</td>
<td>75</td>
<td>65</td>
</tr>
</tbody>
</table>
## Predictions II

<table>
<thead>
<tr>
<th></th>
<th>Left (F1)</th>
<th>Inter (F1)</th>
<th>Right (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All RIGHT</td>
<td>0</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>Regs LEFT</td>
<td>40</td>
<td>0</td>
<td>55</td>
</tr>
<tr>
<td>Basic discr</td>
<td>57</td>
<td>34</td>
<td>53</td>
</tr>
<tr>
<td>Multilevel</td>
<td>60</td>
<td>29</td>
<td>56</td>
</tr>
<tr>
<td>Majority oracle</td>
<td>65</td>
<td>60</td>
<td>70</td>
</tr>
</tbody>
</table>
Conclusions:

- Complex information structure of relational descriptions
- Predictable from visual information...
- More visible objects act like familiar entities

Future work:

- Surface realization of these structures
- More sophisticated visual models