Information structure prediction for visual-world referring expressions

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”
The intersection of:

- Visual perception
  - What do you see?
- Discourse pragmatics
  - What information is useful?
  - Extensive prior work, see (Krahmer+van Deemter ‘12)
- Language production and syntax
  - In what order?
The intersection of:

- Visual perception
  - What do you see?
- Discourse pragmatics
  - What information is useful?
  - Extensive prior work, see (Krahmer+van Deemter ‘12)
- Language production and syntax
  - In what order?
The intersection of:

- **Visual perception**
  - What do you see?
    (Clarke, Elsner and Rohde, in submission)

- **Discourse pragmatics**
  - What information is useful?
  - Extensive prior work, see (Krahmer+van Deemter ‘12)

- **Language production and syntax**
  - In what order?
The intersection of:

- Visual perception
  - What do you see?
    (Clarke, Elsner and Rohde, in submission)
- Discourse pragmatics
  - What information is useful?
  - Extensive prior work, see (Krahmer+van Deemter ‘12)
- Language production and syntax
  - In what order? (this talk)
Setting:

Given abstract structure:

Woman

Jetway

Predict information structure:

The woman standing near the jetway right
An intermediate stage

Information structure:

The woman standing near the jetway

Surface realization:

The woman standing near the jetway
Motivation:

- What factors predict information structure?
- What is the effect of visual features?
Motivation:

- What factors predict information structure?
- What is the effect of visual features?

Theory:

- Structure usually for discourse salience
- We show: visual salience plays similar role

Practice:

- Eventually, text generation...
- With visually aware surface realization
Overview

“Where’s Wally”: the dataset

Ordering strategies in the corpus

Learning to use visual features

Experiments: predicting the order
Where’s Wally

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>`
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
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

right
target1 landmark2

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 right ar, sal, deps dst target1 landmark2 ar, sal, deps

**The woman** standing near **the jetway**
System design

- 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 class: small and hard to see
  - Usually RIGHT of head
  - More dependents (other landmarks) of their own

- Blue class: larger and more visible
  - Often LEFT or INTER
  - Rarer in ESTABLISH construction
  - Few dependents
Information ordered by givenness/familiarity:
(Prince ‘81, Birner and Ward ‘98 etc)

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

Visual salience plays similar role:

- Highly visible landmarks appear on left
  - Assumed in common ground
  - Rarely need their own landmarks
- Harder-to-see landmarks on right
  - Assumed discourse-new
- ESTABLISH construction introduces unfamiliar entity (Ward and Birner ‘95)
How well can we predict 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>
</tbody>
</table>
How well can we predict 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>
How well can we predict 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>
## Results 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:

▶ Does fully hierarchical model work better?
▶ Surface realization of these structures
▶ More sophisticated visual models