Information Structure Prediction for Visual-World Referring Expressions

Micha Elsner
Hannah Rohde, Alasdair Clarke

The Ohio State University
University of Edinburgh
University of Aberdeen
“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
- Leads to complex ordering/coherence preferences
- Image understanding via *visual salience*:
  - Perceptually apparent / attracts attention
- What do they have in common?
- How can we use this in REG?
Ordering strategies: direction

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

**The woman** standing near **the jetway**

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

- Orders defined WRT first mention
- Information structure, not syntax
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
- Almost always occurs with PRECEDE ordering
Basic ordering

- **FOLLOW** (38%) and **PRECEDE** (37%) equally common for landmarks
- **PRECEDE** default for image regions (60%)
  - “On the left of the screen is a woman”...
- **INTER** for 20/25%
- Ordering decisions are non-trivial
This study

- Information ordering for referring expressions is complex
- Visual features matter...
  - Mostly area
- Partly free variation
- Visual salience *is* like discourse salience
Vision affects *content*...

What to say:

(Kelleher et al 05, 06; Duckham 10, Clarke et al 13, Fang et al 13)

- Visual features predict mentioned objects
- Easier to see $\rightarrow$ better landmark
Little work on linguistic form

How to say it:

- Many REG systems only perform content selection (eg Mitchell 12)
- Surface realization for REG: TUNA challenges (Gatt et al 08-10)
  - Standard problems were adjective/phrase orders
  - Templatic approaches were common (Langkilde-Geary, Brugman et al, Di Fabbrizio et al)
- Determiner selection (Duan et al 13)
Where’s Wally: the WREC corpus

Corpus: (Clarke et al 13) Books: (Martin Handford)

- 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
- Available for download!
28 images x 16 targets x 10 subjects per target
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:

- 66% agreement about mention direction
- 43% agree on ESTABLISH constructions

Strategies are predictable but vary

- Based on other landmarks selected?
- Different cognitive strategies?
Predicting the direction

- Construct logistic regression models to predict direction
- Treating each target/landmark pair as independent
- First look at coefficients
- Then accuracies
Features

- Landmark is object or image region?
- Root area of object
- Centrality
- Distance between objects
- Number of landmark objects attached to target
- Scaled to 0 mean and unit var
  - For interpretability
- (Tried visual salience (Torralba ‘06) but didn’t work)
# Coefficients for ordering

| Feature          | PRECEDE | PREC.-EST. | INTER  | FOLLOW |
|------------------|---------|------------|--------|--------|
| intercept        | -4.18   | -2.66      | -2.51  | 2.72   |
| img region?      | 11.46   | -          | 3.01   | -12.62 |

- Image regions strongly prefer to PRECEDE
## Coefficients for ordering

| Feature               | PRECEDE  | PREC. - EST. | INTER  | FOLLOW  |
|-----------------------|----------|--------------|--------|---------|
| intercept             | -4.18    | -2.66        | -2.51 | 2.72    |
| img region?           | 11.46    | -            | 3.01  | -12.62  |
| target area           | -.27     | -.19         | -      | .35     |
| targ centrality       | .11      | -            | -      | -       |
| targ # lmarks         | -        | -.74         | .22   | -       |

- Image regions strongly prefer to PRECEDE
- No strong effects of features of target
| Feature                | PRECEDE | PREC.-EST. | INTER | FOLLOW |
|------------------------|---------|------------|-------|--------|
| intercept              | -4.18   | -2.66      | -2.51 | 2.72   |
| img region?            | 11.46   | -          | 3.01  | -12.62 |
| target area            | -.27    | -.19       | -     | .35    |
| targ centrality        | .11     | -          | -     | -      |
| targ # lmarks          | -       | -.74       | .22   | -      |
| distance               | -       | -.24       | -     | -      |

- Image regions strongly prefer to PRECEDE
- No strong effects of features of target
- No strong effects of distance
Coefficients for ordering

| Feature            | PRECEDE | PREC.-EST. | INTER | FOLLOW |
|--------------------|---------|------------|-------|--------|
| intercept          | -4.18   | -2.66      | -2.51 | 2.72   |
| img region?        | 11.46   | -          | 3.01  | -12.62 |
| target area        | -.27    | -.19       | -     | .35    |
| targ centrality    | .11     | -          | -     | -      |
| targ # lmarks      | -       | -.74       | .22   | -      |
| distance           | -       | -.24       | -     | -      |
| lmark area         | 3.27    | -          | 1.28  | -3.76  |
| lmark centrality   | -       | -          | -     | .81    |

- Image regions strongly prefer to PRECEDE
- No strong effects of features of target
- No strong effects of distance
- Larger landmarks prefer to PRECEDE
Coefficients for ordering

| Feature               | PRECEDE | PREC.-EST. | INTER | FOLLOW |
|-----------------------|---------|------------|-------|--------|
| intercept             | -4.18   | -2.66      | -2.51 | 2.72   |
| img region?           | 11.46   | -          | 3.01  | -12.62 |
| target area           | -.27    | -.19       | -     | .35    |
| targ centrality       | .11     | -          | -     | -      |
| targ # Imarks         | -       | -.74       | .22   | -      |
| distance              | -       | -.24       | -     | -      |
| Imark area            | 3.27    | -          | 1.28  | -3.76  |
| Imark centrality      | -       | -          | -     | .81    |
| Imark # Imarks        | -       | 2.38       | -1.07 | -1.37  |

- Image regions strongly prefer to PRECEDE
- No strong effects of features of target
- No strong effects of distance
- Larger landmarks prefer to PRECEDE
- Landmarks with landmarks prefer own clauses
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)

- Similarly, large landmarks prefer to PRECEDE
Predicting the order

Classification per target/landmark pair

|                | Acc (dir) | F (ESTABLISH) |
|----------------|-----------|----------------|
| **FOLLOW**     | 32        | 0              |
| **PRECEDE**    | 44        | 0              |
| Regions **PRECEDE** | 42  | 0              |
Predicting the order

- Classification per target/landmark pair

|                | Acc (dir) | F (ESTABLISH) |
|----------------|-----------|---------------|
| FOLLOW         | 32        | 0             |
| PRECEDE        | 44        | 0             |
| Regions PRECEDE| 42        | 0             |
| Classifier     | 57        | 60            |
Predicting the order

- Classification per target/landmark pair

|                | Acc (dir) | F (ESTABLISH) |
|----------------|-----------|---------------|
| **FOLLOW**     | 32        | 0             |
| **PRECEDE**    | 44        | 0             |
| Regions **PRECEDE** | 42    | 0             |
| **Classifier** | 57        | 60            |
| **Inter-subject (lbd)** | 66      | 53            |
| **Inter-subject (all)** | 76      | 73            |
Conclusions

For psycholinguists

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

For generation

- Revisit realization for complex descriptions
- Templates may not be sufficient
- Open question: are human-like orders easier to understand?
  - Experiment is in progress...