Summary overview of systems
Summary overview of systems

Algorithms that consider all attribute subsets:

• IS-FBS*: select the minimal AS that has highest Dice similarity with any AS found for referent in training data

• IS-FBN: select the AS of any size that has highest Dice similarity with any AS for referent found in training data

• GRAPH-SC: select AS with lowest sum of attribute costs; cost is derived from attribute frequency relative to all entities with attribute in training set

• GRAPH-FP: same, but some properties are added at zero cost (COLOUR, HASBEARD, HASGLASSES)
Incremental algorithms that do not use discriminatory power:

- IS-IAC: incrementally select individual attributes using a decision tree trained on the training data, backing off to overall most frequent attribute
- DIT: incremental algorithm with attribute order determined offline by absolute frequency in corpus; TYPE always included
  - DIT-DI: use frequency in all of corpus
  - DIT-DS: use frequency in subcorpus (people or furniture)
- NIL: incremental algorithm with empirically determined attribute order for furniture domain, adjusted for people domain
**Summary overview of systems**

**Incremental algorithms that use discriminatory power:**

- **CAM-B:** incremental algorithm with attribute order determined by “discriminating quotient”, for each input; always include TYPE; model HASHAIR/HASBEARD and HAIRCOLOUR dependency
  - CAM-T: as CAM-B, but salience to humans incorporated as weight on discriminating power (frequency-based?)
  - CAM-TU/CAM-BU: update discriminating power at each incremental step
- **TITCH-BS-STAT:** incremental algorithm with attribute order determined by “discrimination power”, for each input; always include TYPE
  - TITCH-AW-STAT: as TITCH-BS-STAT, but discrimination power multiplied by number of times attribute is used in corpus
  - TITCH-RW-STAT: as TITCH-BS-STAT, but discrimination power multiplied by number of times attribute has been found missing compared to corpus data
- **-DYN variants:** discrimination power calculated at each incremental step
- **-PLUS variants:** HASHAIR/HASBEARD and HAIRCOLOUR dependency is additionally modelled
Summary overview of systems

- Both CAM and TITCH systems use notion of discriminatory power $F$ of an attribute-value pair $a$, for a domain $U$ with $N$ entities.

- Originally, Dale (1989): $F(a, U) = (N - n) / (N - 1)$, where $n$ is the number of entities in the domain for which $a$ is true.

- **CAM:** $F(a, U) = (N - n) - (n - 1)$

- **TITCH:** $F(a, U) = N - n$
Properties of algorithms

- Non-incremental/incremental, with offline/online attribute ordering, updated at each step – yes/no
- Consider distractors and their properties (discriminatory power) – yes/no
- Non-trainable/trainable
- Hardwire inclusion of type – yes/no
- Hardwire dependency between HAIRCOLOUR and HASBEARD/HASHAIR – yes/no
Evaluation results
Evaluation criteria

1. Uniqueness
2. Minimality
3. Humanlikeness
4. Identification Speed
5. Identification Accuracy
Criterion 1: Uniqueness

• Method:
  • for each output, we determined set of all matching referents
  • then computed percentage of outputs for which size of set = 1

• Results: all systems except one scored 100%
• Exception: TITCH-AW-DYNAMIC which scored 77%
Criterion 2: Minimality

- Method:
  - for each output, we determined size of minimal set for the referent, and whether output was the same size (not checking for uniqueness)
  - then computed percentage of outputs of minimal size
- Results: 22 systems range from 14.86% to 93.92% minimal outputs
Minimality results

TITCH-AW-DYN
TITCH-BS-STAT
CAM-BU
CAM-B
GRAPH-SC
TITCH-RW-DYN
TITCH-RW-STAT
TITCH-AW-DYN-PLUS
IS-FBS
GRAPH-FP
CAM-T
CAM-TU
NIL
DIT-DS
IS-JAC
IS-FBN
DIT-DI
Minimality results

Non-trainable systems
Minimality results

Non-incremental systems
Minimality results

No discriminatory power
Criterion 3: Humanlikeness

• Method: compute Dice coefficient between system outputs $A_1$ and (human-produced) reference attribute sets $A_2$

$$\frac{2 \times (|A_1 \cap A_2|)}{|A_1 \cup A_2|}$$

• Results:
  • Self-reported Dice scores on development data
  • Dice scores computed by organisers on test data
  • Correlation: Pearson’s $r = 0.93$, $p < 0.001$
Humanlikeness – development set results
Humanlikeness – test set results

Overall, furniture, and people categories show varying levels of humanlikeness across different models.
Humanlikeness – development set vs. test set results
Humanlikeness – test set results

Non-trainable systems
Humanlikeness – test set results

Non-incremental systems
Humanlikeness – test set results

No discriminatory power
Humanlikeness – test set results

Not always include TYPE
Humanlikeness – test set results

Not model HAIR dependency
Minimality scores in order of Dice

- IS-FBN
- DIT-DS
- IS-JAC
- CAM-T
- CAM-TU
- GRAPH-SC
- TITCH-RW-STAT
- TITCH-RW-DYN
- TITCH-RW-DYN-PLUS
- TITCH-BS-DYN
- TITCH-BS-STA
- TITCH-BS-STA-PLUS
- CAM-BU
- NIL
- DIT-DI
- CAM-B
- TITCH-BS-DYN
- TITCH-BS-STA
- IS-FBS
Humanlikeness – statistical significance

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IS-FBN</td>
<td>0.7709</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIT-DS</td>
<td>0.7501</td>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>IS-IAC</td>
<td>0.7461</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>CAM-T</td>
<td>0.7249</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>CAM-TU</td>
<td>0.7214</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>GRAPH-FP</td>
<td>0.6898</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>GRAPH-SC</td>
<td>0.6715</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>TITCH-RW-STAT</td>
<td>0.6551</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>TITCH-RW-DYN</td>
<td>0.6551</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>TITCH-AW-STAT-PLUS</td>
<td>0.6532</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>TITCH-AW-DYN-PLUS</td>
<td>0.6532</td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>TITCH-AW-STAT</td>
<td>0.6455</td>
<td></td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>TITCH-AW-DYN</td>
<td>0.6411</td>
<td></td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>TITCH-RW-STAT-PLUS</td>
<td>0.6400</td>
<td></td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>TITCH-RW-DYN-PLUS</td>
<td>0.6400</td>
<td></td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>CAM-BU</td>
<td>0.6300</td>
<td></td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>NIL</td>
<td>0.6251</td>
<td></td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>DIT-DI</td>
<td>0.6243</td>
<td></td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>CAM-B</td>
<td>0.6203</td>
<td></td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>TITCH-BS-DYN</td>
<td>0.5934</td>
<td></td>
<td>E</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>TITCH-BS-STAT</td>
<td>0.5928</td>
<td></td>
<td>E</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>IS-FBS</td>
<td>0.5276</td>
<td></td>
<td>F</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Task-based evaluations: Identification Speed and Accuracy

- Outputs of 15 systems submitted by participants were evaluated in an identification experiment
- Repeated Latin Squares Design
- 30 subjects, experienced computer/mouse users, no NLP background
- Simple template realiser (created by Irene Langkilde-Geary) used to turn attribute sets into REs
- Subjects shown RE and pictures representing domain entities
- Total of 2,250 trials
- Used DMDX and TimeDX (Forster and Forster, 2003) to display text/pictures, and measure timings (millisecond accuracy)
- Recorded for each trial:
  - Whether or not intended referent was selected (Criterion 4: Identification Accuracy)
  - Time taken by subject from display to identification (Criterion 5: Identification Speed)
the red chair facing right
Criterion 4: Identification accuracy

- Method: for each system, compute proportion of times wrong referent is selected
- Results: no significant differences between systems; accuracy ranges from 85% to 91%
Criterion 5: Identification speed

- Method: for each system, compute average time between RE appearing on screen and subject selecting picture
- Results: there were significant differences between systems; three homogeneous subsets
<table>
<thead>
<tr>
<th>Identification speed – results table</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITCH-RW-STAT</td>
</tr>
<tr>
<td>CAM-TU</td>
</tr>
<tr>
<td>CAM-T</td>
</tr>
<tr>
<td>TITCH-AW-STAT-PLUS</td>
</tr>
<tr>
<td>CAM-BU</td>
</tr>
<tr>
<td>GRAPH-FP</td>
</tr>
<tr>
<td>TITCH-RW-STAT-PLUS</td>
</tr>
<tr>
<td>CAM-B</td>
</tr>
<tr>
<td>DIT-DS</td>
</tr>
<tr>
<td>GRAPH-SC</td>
</tr>
<tr>
<td>IS-IAC</td>
</tr>
<tr>
<td>TITCH-AW-STAT</td>
</tr>
<tr>
<td>NIL</td>
</tr>
<tr>
<td>IS-FBN</td>
</tr>
<tr>
<td>IS-FBS</td>
</tr>
</tbody>
</table>
Minimality in order of identification speed

- TITCH-RW-STAT
- CAM-TU
- CAM-T
- CAM-BU
- CAM-B
- DIT-DS
- GRAPH-SC
- IS:IC
- TITCH-AW-STAT
- N!L
- IS:FBN
- IS:FBS
- TITCH-AW-DYN
- TITCH-RW-STAT
- TITCH-RW-DYN
- TITCH-RW-DYN-PLUS
- DIT-DI
Conclusions

• Results of the ASGRE Challenge do not tell us what the best way to do GRE is
• Rather: results for 22 systems and 5 quality criteria which can help guide development and choice of GRE methods (in similar domains), especially when aiming to maximise specific criteria
• Tentative generalisations:
  • Trainable systems generally scored higher on Dice
  • Some evidence that Dice and minimality are negatively correlated
  • Some evidence that minimality and identification time are negatively correlated