LEARNING TO SPORTSCAST: A TEST OF GROUNDED LANGUAGE ACQUISITION

Rose Zhang
Children learn language by connecting it to their environment.

Machine learning would ideally be able to learn by “grounding” language to visual input as well.

As step in that direction, we can use a simulated soccer game to generate commentary by learning how human commentaries are paired with events in the game.
ROBOCUP SIMULATOR

Uses a simulated game of robot soccer with written commentary

- Circumvents computer vision and speech to text issues to focus on language learning
Certain soccer events (kicking, passing, offside, corner kick, etc.) are extracted and represented by predicate logic. These logical facts = meaning representations (MRs).

Human commented simulation games from Robocup

Each comment is paired with all events from 5secs or less of the comment

Gold standard created from manually matching each comment with the correct MR if it exists (bold lines)
ALGORITHMS

WASPER

- Extension of WASP which uses modified statistical machine translation (SMT) to translate NL sentences to MRs
- Same learned model can also generate sentences from MRs
- WASP needs unambiguous training data, but one comment can have multiple events/MRs
- WASPER uses retraining to deal with ambiguity

KRISPER-WASP

- KRISP: SVMs and string kernels to construct a set of probable MRs for sentence. Then retrain to improve selection of unambiguous training set (part of KRISPER)
  - Can't generate sentences from MRs
- Then use set to train WASP for generation
RETRAINING: FIRST ITERATION

Purple goalie turns the ball over to Pink8

Pink8 passes to Pink11

Pink8 passes back to Pink11

Turnover(PurplePlayer1, PinkPlayer8)

Kick(PinkPlayer11)

Kick(PinkPlayer8, PinkPlayer11)
Purple goalie turns the ball over to Pink8

Pink8 passes to Pink11

Pink8 passes back to Pink11

WASP/KRISP

Turnover(PurplePlayer1, PinkPlayer8)

Kick(PinkPlayer11)

Kick(PinkPlayer8)

Pink8 passes back to Pink11
RETRAINING

Natural Language Commentary

Purple goalie turns the ball over to Pink8

Pink8 passes to Pink11

Pink8 passes back to Pink11

Meaning Representation

badPass (PurplePlayer1, PinkPlayer8)
turnover (PurplePlayer1, PinkPlayer8)

pass (PinkPlayer8, PinkPlayer11)
kick (PinkPlayer11)

pass (PinkPlayer11, PinkPlayer8)
kick (PinkPlayer8)

Pass(PinkPlayer8, PinkPlayer11)

Kick(PinkPlayer8)

WASP/KRISP
Purple goalie turns the ball over to Pink8
Pass(PinkPlayer8, PinkPlayer11)
Pass(PinkPlayer8, PinkPlayer11)
Turnover(PurplePlayer1, PinkPlayer8)
Pass(PinkPlayer8, PinkPlayer11)
Pass(PinkPlayer8, PinkPlayer11)
WASP/KRISP
Instead of parsing sentence to MRs, evaluate likelihood of generating sentence given MR.

For NL-MR pair \((n, m)\), generate a sentence for \(m\), and score generated sentence against \(n\)

- NIST score: precision of translation in terms of proportion of n-grams it shares with human translations
Not just about how to say something, but also what to say
- Given multiple events at same time, what do we commentate on?

For each event type, predict probability that sportscaster will announce it
- Easy with gold standard matching of NL-MR, but system doesn’t know correct match and must estimate

Percentage of an event type that’s matched to a sentence = probability of commenting (from previous algorithms)

Iterative Generation Strategy Learning (IGSL)
- Initially, each MR to sentence match = 1/# of MRs in sentence
- Uses iterative training to improve estimates
EVALUATION

F-measure: harmonic mean of recall and precision
- Precision: fraction of system’s annotations that are correct
- Recall: fraction of gold standard annotations correctly produced

NIST scores: roughly estimates how well produced sentences match target sentences
MATCHING NL AND MR

How well do various systems (WASPER, KRISPER, etc.) pick the correct NL-MR pair?

WASPER only does better than random matching

KRISPER expected to do better since it handles noisy data better, but it ties with WASPER-GEN
SEMANTIC PARSING

Produce MRs for each human written sentence

Each system produces an MR for each sentence in the test set that has a gold standard matching

Parse is correct iff it matches the gold standard exactly

Difficult because different ways to describe the same thing (kick vs pass, player1 vs goalie)

Results follow matching
GENERATION

All systems are given an MR in the test set that was in the gold standard and told to generate a sentence from it.

Quality measured by comparing to gold standard using NIST

Easier than parsing: only need to learn one way

WASPER and KRISPER-WASP do the worst as systematic errors have greater consequences

WASPER-GEN does the best
STRATEGIC GENERATION

IGSL does best besides gold standard

Everything else is inferred based on percentage of an event type that is matched to a sentence

Inferred does worse on rare but significant events like goals
Automatic evaluation not quite as good as human evaluation

4 fluent English speakers as judges

Fluency = grammar and syntax

Correctness = commentary accuracy

Sportscasting = interesting and flows well

<table>
<thead>
<tr>
<th>Score</th>
<th>English Fluency</th>
<th>Semantic Correctness</th>
<th>Sportscasting Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Flawless</td>
<td>Always</td>
<td>Excellent</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Usually</td>
<td>Good</td>
</tr>
<tr>
<td>3</td>
<td>Non-native</td>
<td>Sometimes</td>
<td>Average</td>
</tr>
<tr>
<td>2</td>
<td>Disfluent</td>
<td>Rarely</td>
<td>Bad</td>
</tr>
<tr>
<td>1</td>
<td>Gibberish</td>
<td>Never</td>
<td>Terrible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>English Fluency</th>
<th>Semantic Correctness</th>
<th>Sportscasting Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>3.938</td>
<td>4.25</td>
<td>3.625</td>
</tr>
<tr>
<td>Machine</td>
<td>3.438</td>
<td>3.563</td>
<td>2.938</td>
</tr>
</tbody>
</table>

**HUMAN EVALUATION**
Critique

Comments are matched to events that occur up to 5 seconds before the comment. Not all comments refer to events within 5 secs of commenting.

Since there are different ways to describe the same thing, could there be results that don’t exactly match the gold standard but are valid?

Limited meaning representation language and how to account for events that can’t be translated easily to MR (Player is sloppy)

How were the four “judges” chosen and do they have any bias that would show especially given the small sample size?
FUTURE RESEARCH

Expand MRL to account for temporal and spatial events as well as more opinion based statements vs factual events

Look into refining the time limit to attach events to sentences

Apply model to generate captions on video by integrating computer vision
QUESTIONS?