Shallow Semantic Reasoning from an Incomplete Gold Standard for Learner Language

Levi King, Markus Dickinson
Indiana University

Overview

- **Goal:** Evaluate semantic accuracy of non-native speaker (NNS) responses to picture description task (PDT)
  - Compare to gold standard (GS) of native speaker (NS) responses

Past Approach

1. Dependency parse NNS responses & GS
2. Use custom rules to extract and lemmatize *verb-subject-object* triple for each response
3. Attempt to match NNS triple to GS triples

Past Limitations

1. GS is small
2. NNS responses show more variation than NS responses
3. Matching exact triples is restrictive (no partial matching)
   - *kick(boy, ball)* ≠ *kick(boy, football)*

Upshot: low coverage (50.8%)

Current Approach

Generalize methods by:

1. Representing responses as lists of dependencies
2. Scoring NNS response representation according to how closely it resembles GS representation
   - partial matching: **subj_boy_kick** + **obj_ball_kick** vs. **subj_boy_kick** + **obj_football_kick**

Data

Picture Description Task (PDT)

- 10 items (2 photos, 8 drawings) depicting transitive events
- PDT elicits natural productions but constrains form & content

Participants

- 39 NNSs, intermediate/advanced English; 390 sentences
  - Arabic, Chinese, Japanese, Korean, Spanish, Kurdish, Polish, Portugese
- 14 NSs; 140 sentences

Example Response (L1)
The man killing the beard. (Arabic)
A man is shutting a bird. (Chinese)
A man is shooting a bird. (English)
The man shouted the bird. (Spanish)

Generalizing the Methods

In the sections below, we explain the system parameter settings. The first two are closely related to generalizing the methods to overcome a limited GS, handle a wider range of sentence types (beyond transitives), and better reflect similarity to the GS.

- Response Representation: By moving to a "bag of dependencies" approach, we loosen the strict evaluation from *covered/not covered*; partial dependencies further loosen matching.
- Response Scoring: Averaging response term scores or calculating cosine distances allows for gradable rather than binary response scoring.

Response Representation

Responses are dependency parsed and treated as a list of *terms*, which are dependencies in one of the formats below (*l, d, h = label, dependent, head; x = placeholder*):

- **ldh:** `subj_boy_kick`
- **xdh:** `x_boy_kick`
- **lxh:** `subj_x_kick`
- **ldx:** `subj_boy_x`
- **xdx:** `x_boy_x`

Response Scoring

Scoring responses involves:

- Weighting terms (dependencies)
- Scoring responses: comparing weighted NNS terms with weighted GS terms

- **Frequency Average (FA):**
  - Weight: NNS terms assigned GS term frequencies
  - Response score: average of NNS term scores
- **Tf-idf Average (TA):**
  - Weight: NNS terms assigned tf-idf scores based on GS frequencies
  - Response score: average of NNS term scores
- **Frequency Cosine (FC):**
  - Weight: NNS & GS term frequencies are calculated
  - Response score: cosine distance between NNS & GS term scores
- **Tf-idf Cosine (TC):**
  - Weighting: NNS & GS tf-idf values are calculated
  - Response score: cosine distance between NNS & GS term scores

Reference Corpus

`TA` and `TC` require a reference corpus for deriving tf-idf scores. We experimented with two:

- **Brown Corpus (B)**
- **Wall Street Journal Corpus (W)**

NNS Source

We experiment with two forms of the NNS responses:

- **NNSO:** Original, uncorrected form
- **NNSLM:** Language Model autocorrected form

System Evaluation

1. Manually annotate responses; unacceptable responses are "errors"
2. Use each combination of parameters to produce a scored, ranked list of responses (Table 1)
   - Good parameter settings should rank good responses near GS and errors far from GS
3. Evaluate and rank parameter settings by (mean) average precision ((M)AP) (Table 3)
   - Also evaluate settings by non-normalized error score, which better illustrates differences in difficulty of PDT items (used in Figure 2)
4. Evaluate individual parameter values by MAP (Table 2)

R	*S*	Sentence	*E*	*V*
1	1.000	she is hurting.	1	1.5
	1.000	man mull bird	1	1.5
3	0.996	the man is hurting duck.	1	3.0
4	0.990	he is hurting the bird.	1	3.0
11	0.865	the man is trying to hurt a bird	1	11.0
12	0.856	a man hunted a bird.	0	0.0
17	0.775	the bird not shot dead.	1	17.0
18	0.706	he shot at the bird	0	0.0
19	0.669	a bird is shot by a un	1	19.0
20	0.646	the old man shooting the birds	0	0.0
37	0.086	the old man shot a bird.	0	0.0
38	0.084	a old man shot a bird.	0	0.0
39	0.058	a man shot a bird	0	0.0
		Total Raw Score (not normalized)	17	169
		Average Precision	0.75084	

Table 1: Excerpt of rankings for Item 10 from the best system setting (TC_B_NNSLM_ldh) based on average precision scores. *R*: rank; *S*: sentence score; *E*: error; *V*: rank value.

Results

Approach		Term Form		Ref. Corp. (TA/TC)		NNS Source	
0.51577	TC	xdh	0.51810	Brown	0.51534	NNSLM	0.51937
0.50780	FC	ldh	0.51677	WSJ	0.50798	NNSO	0.49699
0.50755	TA	lxh	0.51350				
0.49464	FA	xdx	0.49901				
		ldx	0.49352				

Table 2: Approaches and parameters ranked by mean average precision for all 10 PDT items.

- Best approach: **TC**
  - **TC** > **FC**, **TA** > **FA**:
    - tf-idf weighting > frequency weighting
  - **TC&FC** > **TA&FA**:
    - cosine distance > weight averaging
- Term form: **xdh, ldh, lxh** > **xdx, ldx**
  - Importance of heads (**h**): with short transitive responses, verbs are salient (subj/obj head)
- Reference corpus: **Brown** > **WSJ**
  - Content & style of responses more like **Brown**
- NNS source: **NNSLM** > **NNSO**
  - More errors in NNSLM forms, inflating MAP values: use non-normalized scores? (see paper)

Rank	MAP	Settings
1	0.5534	TC_B_NNSLM_lxh
2	0.5445	TA_B_NNSLM_lxh
3	0.5435	TC_W_NNSLM_lxh
4	0.5422	TC_B_NNSLM_xdh
5	0.5368	TC_B_NNSLM_ldh
56	0.4816	TA_B_NNSO_xdx
57	0.4796	FA_na_NNSLM_ldx
58	0.4769	FC_na_NNSO_lxh
59	0.4721	TA_W_NNSO_xdx
60	0.4530	FA_na_NNSO_lxh

Table 3: Based on Mean Average Precision, the five best and five worst settings across all 10 PDT items.

Future Directions: Clustering Items By Features

We used hierarchical clustering to explore for patterns among the items and parameters.
Set-up: cluster PDT items using features from response (e.g., type/token counts for terms) & features from system performance (i.e., average error score for parameter setting).
Goal: new PDT items could be placed into known clusters via response features & optimal parameter settings for that cluster could be applied automatically

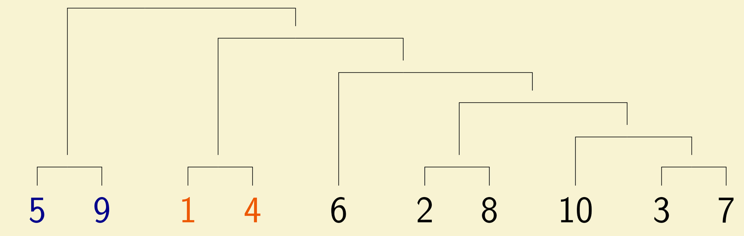

Figure 1: PDT items clustered by type and token counts of all NS, NNSO and NNSLM responses.

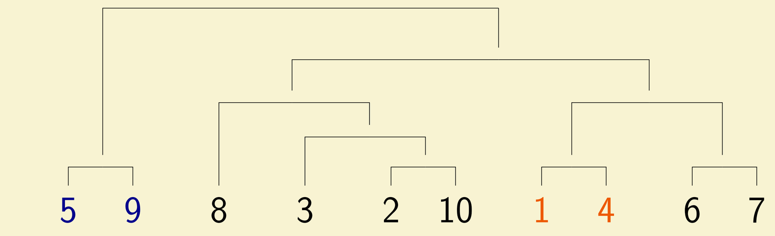

Figure 2: PDT items clustered by parameter performance.

{leviking,md7}@indiana.edu