Semantic Analysis of Image-Based Learner Sentences

Levi King

Dissertation Proposal

Department of Linguistics, Indiana University

August 29, 2016

Motivation

Issue:

- Intelligent Computer-Assisted Language Learning (ICALL) / Intelligent Language Tutor (ILT) systems tend to focus on grammatical errors & feedback (Heift and Schulze, 2007; Meurers, 2012).
- Second Language Acquisition (SLA) research has established:
  - Explicit grammar instruction & feedback are often ineffective (Ellis, 2006).

Goals:

- Big Picture: Steer ICALL/ILT toward a focus on communication, with learners producing *more* target language rather than perfectly formed target language.
  - Requires better methods to provide semantic analysis of contextual learner sentences.
- Current: Explore the viability of such analysis using existing tools and resources.

Motivation

Previous Work

- *Herr Komissar*: ILT/detective game for German learners; content analysis & sentence generation (DeSmedt, 1995), but uses many custom tools.
- Petersen (2010): ILT, provides feedback on questions in English, extracting meanings from an existing parser.
- Content assessment: (e.g., ETS's c-rater system (Leacock and Chodorow, 2003)); mostly focused on essay & short answer scoring.
  - Some focus on semantic analysis under restricted conditions, e.g., (Meurers et al., 2011).
- Somasundaran and Chodorow (2014), Somasundaran et al. (2015): score picture-based sentences & narrations
  - for relevance: calculate overlap of response contents with picture contents
  - for picture contents: expert annotators listed all items in the picture & wrote 5 sentences each

Motivation

Reasoning about learner meaning

Focusing on semantic analysis (feedback, scoring, etc.) means that one should have a sense of:

- Semantic correctness / relevance of learner sentences
  - ... vs. appropriateness / nativeness
- Semantic variability in learner data

Still teasing these apart; forming annotation scheme

Note that this is semantic analysis given a gold standard (GS) of native sentences

- Image description often uses semantic primitives (Gilberto Mateos Ortiz et al., 2015)
- For learner data, we want to ensure that we can account not just for correct semantics (*what*), but natural expressions of it (*how*)
  - i.e., we need access to specific linguistic forms (GS)

Background

General approach

I approximate the goal of an ICALL application that evaluates semantic accuracy and appropriateness by:

1. collecting data from a task which elicits contextual language use, namely a picture description task (PDT),
2. parsing it with an off-the-shelf parser,
3. comparing NNS response dependencies with GS set of NS dependencies to get similarity score,
4. ranking NNS responses by score,
5. discriminating between acceptable and unacceptable responses based on rankings (*work in progress*)

Background

Data

Response (L1)
He is droning his wife pitcher. (Arabic)
The artist is drawing a pretty women. (Chinese)
The artist is painting a portrait of a lady. (English)
The painter is painting a woman's paint. (Spanish)

- 10 items, mostly transitives
- 14 NSs, 39 NNSs

Research questions

Overview

1. Are NSs & NNSs similar enough for auto analysis?
2. How to represent responses internally?
3. Can existing tools & resources be used effectively? Which ones?
4. Effectiveness of bag-of-words vs. bag-of-dependencies?
5. Can semantic tools improve performance?
6. Annotation scheme? System vs. human agreement?

Research questions

1. Are the responses of intermediate and advanced L2 English learners sufficiently similar to those of NSs to allow automatic evaluation based on a collection of NS responses? In other words, do learners demonstrate significant overlap with native-like usage in a PDT setting?

Research questions

1. Are the responses of intermediate and advanced L2 English learners sufficiently similar to those of NSs to allow automatic evaluation based on a collection of NS responses? In other words, do learners demonstrate significant overlap with native-like usage in a PDT setting?

2. In the constrained communicative environment of a PDT, what are appropriate response and GS representations for the purpose of providing meaning-oriented feedback or evaluation? In other words, which linguistic components are crucial and which are superfluous?

Research questions

3. What kinds of existing NLP tools and language resources can be integrated to form a content analysis system for open response language learning tasks?

Research questions

3. What kinds of existing NLP tools and language resources can be integrated to form a content analysis system for open response language learning tasks?

4. How do "bag-of-words" and "bag-of-dependencies" approaches compare in terms of performance? Is a bag-of-words approach alone adequate for this task?

Research questions

5. Can the accuracy of the system be improved by the inclusion of semantic information from tools like semantic role labelers or WordNet?

Research questions

5. Can the accuracy of the system be improved by the inclusion of semantic information from tools like semantic role labelers or WordNet?

6. What is the annotation scheme for this task and can the system perform within the range of inter-annotator reliability? Relatedly, what does it mean for a response to be *appropriate* and how can this be captured with annotation?

Past work

Results

King and Dickinson (2013) (no spelling correction):

- matched NNS S-V-O with NS S-V-O set (GS)
- 92-93% extraction accuracy (extracted correct S-V-O)
- Among correctly extracted triples, we achieved 50.8% coverage (153/301), 60.6% accuracy (218/360)

King and Dickinson (2014) (with spelling correction):

- Decreased errors by 13.7% (*with potential for a decrease of 25.9%)
  - *Spelling correction is influenced by the GS, which is very limited here
- Boosted coverage by 13.4% (again, influenced by limited GS)

Past work

Limitations

There are several limitations to this past approach:

- Incomplete data:
  - The GS will always be missing innovations of learners
- Incomplete rules
  - Extraction rules do not generalize beyond transitives
- Simplistic notion of non-native speaker data:
  - Responses are not always simply right or wrong, but perhaps more or less relevant, nativelike, ...

Dissertation work: Addressing these limitations!

Dissertation work

Remaining tasks

- Data collection: 25 items; more forms, participants
- Annotation: scheme; guidelines; implementation [1][6]
  - inaccurate $<$ accurate, not native-like $<$ accurate + native-like
- Revise system to match new annotation scheme [2,6]
- Run system on new data [1,6]
- Integrate semantic role labeler & WordNet; get results [3,4,5]
- Clustering & automatic parameter selection; new results [1,6]
- Feedback module: provide user with most-similar NS response, most common NS response [2]

[1] "[ ]": Relevant research questions

Dissertation work

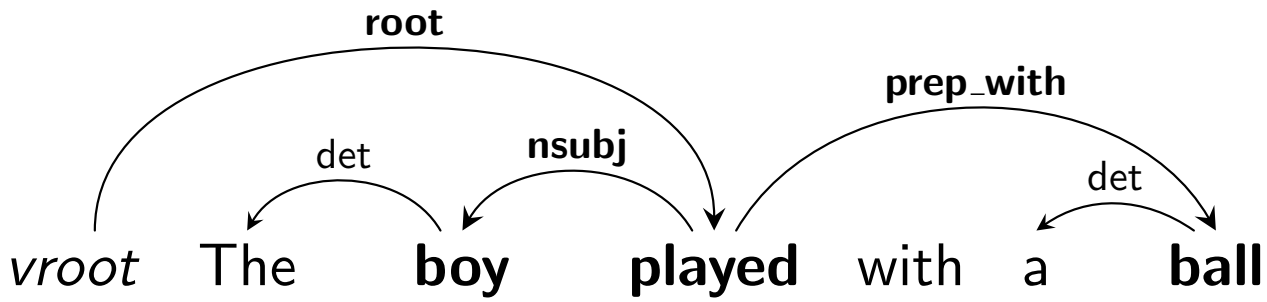
Response representation

Approach: Generalize representations to overcome incomplete GS by using dependencies

- Simpler than triples: no extraction rules or errors
- Can distribute matching over smaller, overlapping pieces of info, not *single, highly specific* piece of info (SVO)

Currently: List out all dependencies in response

- Full form (label#dependent#head): `nsubj#boy#play`
- Plus partial abstractions: `nsubj#X#play`, etc.



Dissertation work

Response and GS comparison

Approach: Loosen the exact matching requirements & use frequency of terms to determine importance

- i.e., deduce relevancy from consistency among NSs

Four (2×2) comparison approaches:

- Term (i.e., lemmatized dependency or word) scores:
  1. frequency; *vs*
  2. tf-idf [2]
- Response score:
  1. assign NNS terms their scores in GS, then average; *vs*
  2. compare NNS & GS term score vectors for cosine distance

[2] term frequency-inverse document frequency; method for scoring importance of terms in a document by comparing frequencies in document to those in a general sample of the language

Dissertation work

Experiments

Rather than *match* / *non-match* determination, current approaches give response scores ranging from 0–1.

Parameters:

1. term form (label/dep/head): ldh; *x*dh; l*x*h; ld*x*
   - *OR* individual words (lemmatized)
2. tf-idf reference corpus: Brown; WSJ
3. NNS source: Original (NNSO); Version from spelling correction + LM (NNSLM)
   - Also planning joint analysis: both sources are considered and the one with the better score is selected.

Dissertation work

Ranking sentences

Each set of parameters produces a scored, ranked list of responses.

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	

R: rank; S: sentence score; E: error; V: rank value. Item 10, best system setting (TC_B_NNSLM_ldh) based on average precision scores.

Dissertation work

Data collection

- 25 items; various forms, syntax
- NSs provide 2 responses per item

intransitive?

ditransitive?

passive?

compound subj?

Future work

Timeline

Task	Start	Finish
System (incl. SRL/WordNet, clustering)	...	Dec 2016
Write introduction	Aug 2016	Sep 2016
Write SLA section (motivation, lit review)	Aug 2016	Oct 2016
Prep PDT; IRB; collect data	Jun 2016	Oct 2016
Write NLP section (lit review)	Sep 2016	Nov 2016
Write method section (incl. training data)	Jul 2016	Dec 2016
Annotation	Sep 2016	Nov 2016
Write data collection & description section	Sep 2016	Nov 2016
System experiments with new data	Nov 2016	Dec 2016
Write experiments & results section	Dec 2016	Jan 2017
Implement feedback system	Jan 2017	Feb 2017
Write feedback section	Jan 2017	Feb 2017
Write conclusion (finish 1st draft)	Jan 2017	Feb 2017
Complete any follow-up experiments	Feb 2017	Mar 2017
Revise draft	Mar 2017	Apr 2017
Final draft complete / Defense		May 2017

References

William DeSmedt. 1995. Herr Kommissar: An ICALL conversation simulator for intermediate German. In V. Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology*, pages 153–174. Lawrence Erlbaum, Mahwah, NJ.

Rod Ellis. 2006. Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40:83–107.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515. Association for Computational Linguistics, Denver, Colorado.

Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Levi King and Markus Dickinson. 2013. Shallow semantic analysis of interactive learner sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–21. Atlanta, Georgia.

Levi King and Markus Dickinson. 2014. Leveraging known semantics for spelling correction. In *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, pages 43–58. Uppsala, Sweden.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and Humanities*, pages 389–405.

Detmar Meurers. 2012. Natural language processing and language learning. In Carol A. Chapelle, editor, *Encyclopedia of Applied Linguistics*. Blackwell.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *Special Issue on Free-text Automatic Evaluation. International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL)*, 21(4):355–369.

Kenneth A. Petersen. 2010. *Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter?* Ph.D. thesis, Georgetown University, Washington, DC.

Swapna Somasundaran and Martin Chodorow. 2014. Automated measures of specific vocabulary knowledge from constructed responses ('Use these words to write a sentence based on this picture'). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11. Baltimore, Maryland.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48. Association for Computational Linguistics, Denver, Colorado.

Response and GS comparison

The 2×2 dimensions result in these approaches:

- **FA** (freq avg): Baseline.
  - Each NNS term is scored with the term's frequency in the GS. Response term scores are averaged.
  - Higher scores are closer to GS.
- **TA** (tf-idf avg): Run tf-idf on GS.
  - Each NNS term is scored with the term's tf-idf score in GS. Response term scores are averaged.
  - Higher scores are closer to GS.
- **FC** (freq compar.): Calculate term freq for NNS, GS; Treat as vectors, score response with cosine distance.
- **TC** (tf-idf compar.): Run tf-idf on NNS, GS; Treat these as vectors, score response with cosine distance.

Comparison results

Evaluating settings

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

Comparison results

Preliminary observations

- 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?

Comparison results

Evaluating settings

Ranked approach & parameter scores (averaged across all 10 PDT items)

Approach		Term Form		tf-idf reference		NNS Source	
TC	0.51577	xdh	0.51810	Brown	0.51534	LM	0.51937
FC	0.50780	ldh	0.51677	WSJ	0.50798	Orig	0.49699
TA	0.50755	lxh	0.51350				
FA	0.49464	xdx	0.49901				
		ldx	0.49352				

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

Clustering

Feature selection

Currently, experimenting with combinations of features for hierarchical clustering of PDT items. (And settings?)

- To help get a handle on learner variability for different items and correlation to best-performing models

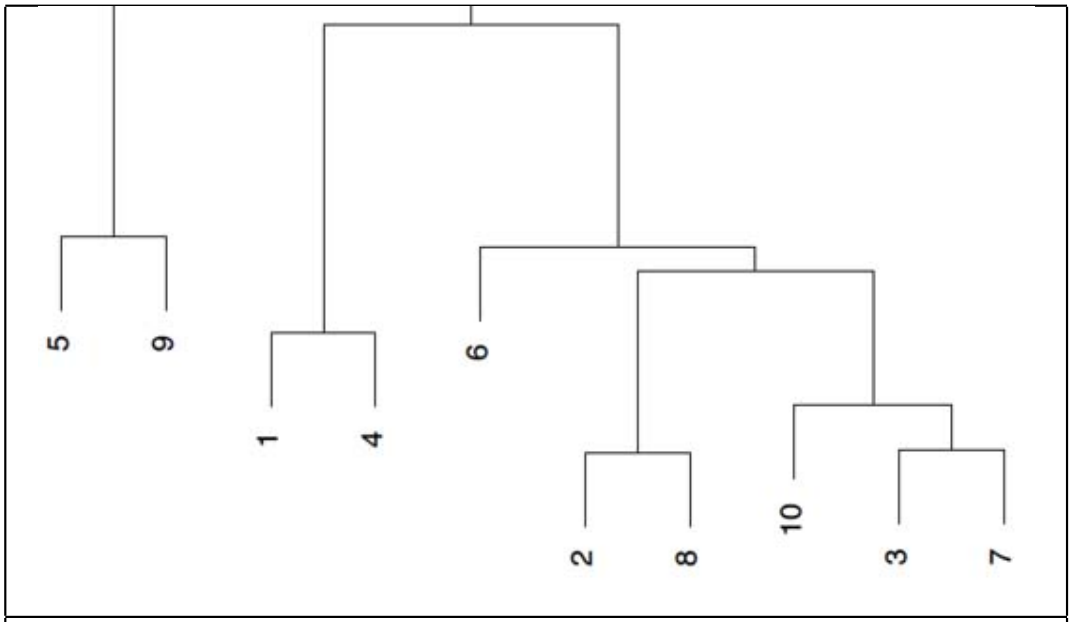
Response features:

- For each response source (GS, NNSO, NNSLM):
  - For each term form (lemma, ldh, xdh, lxh, ldx):
  - (We also use previous triples as "terms" here)
    - type count
    - token count
    - type-to-token ratio

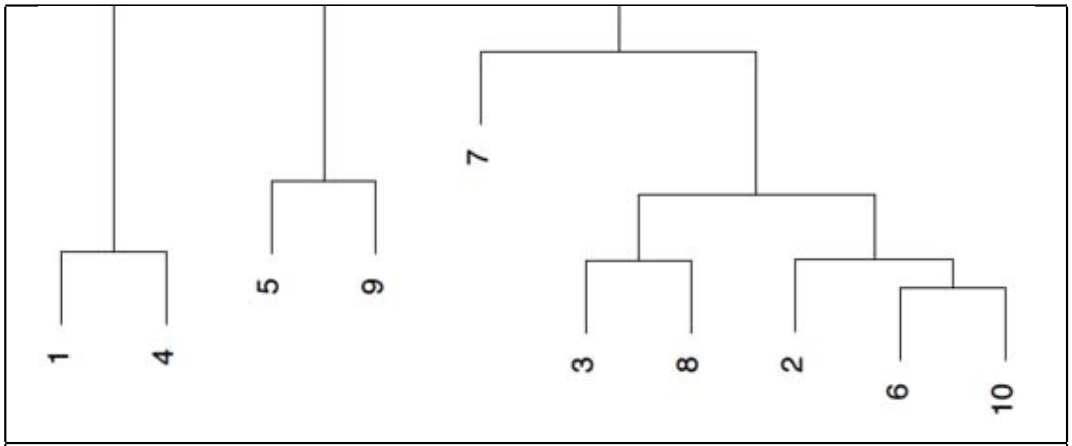
System features:

- Per item average error scores for settings using:
  - Each approach (FA, TA, FC, TC)
  - Each tf-idf reference (Brown, WSJ)
  - Each NNS response source (NNSO, NNSLM)
  - Each term form (lemma, ldh, xdh, lxh, ldx)

Clustering items by responses

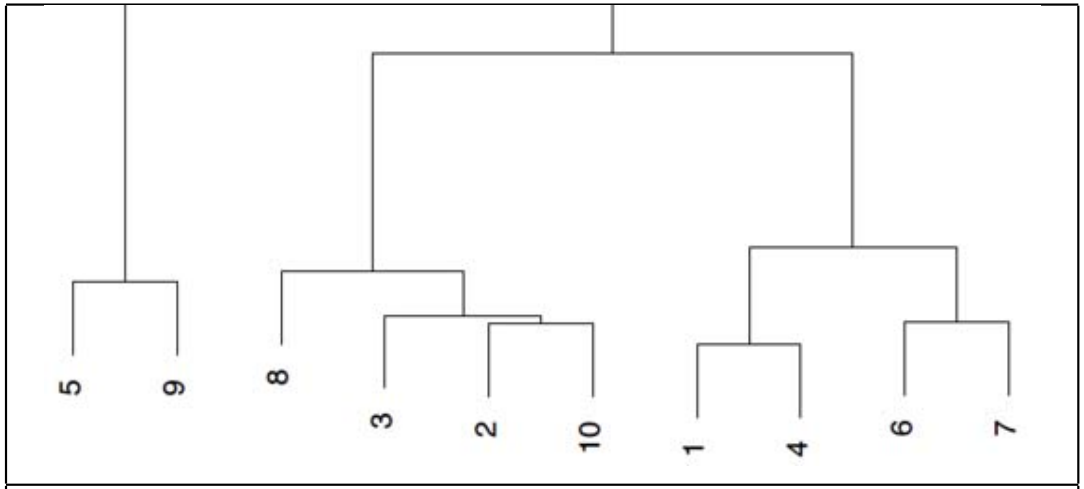


All NNSO, NNSLM, GS type & token counts

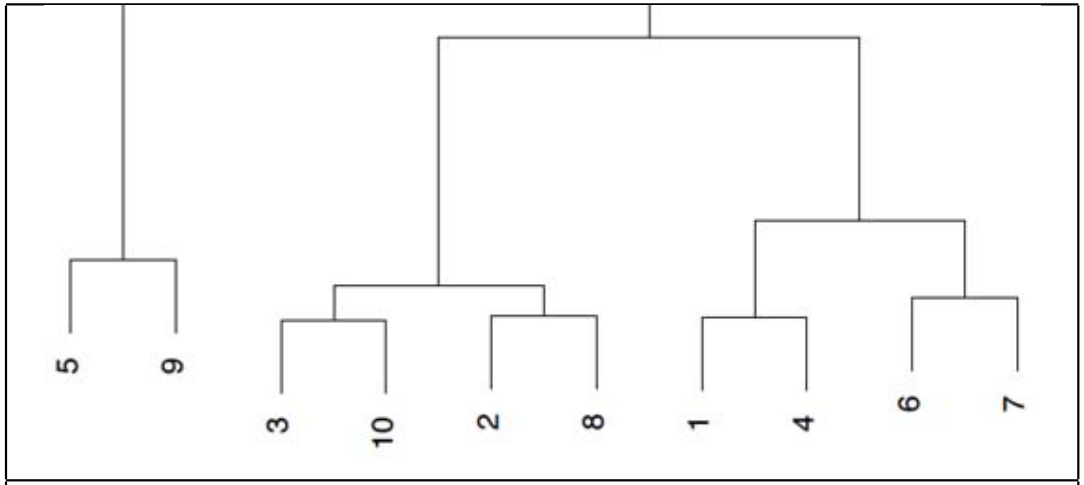


NNSO type-to-token ratio

Clustering items by system performance



All average scores for approaches & parameters



All average scores for approaches

Clustering

Observations

Some patterns among clusters:

	Sentence	Observations
5	A man is raking leaves.	TA & TC are high
9	Two boys are rowing a boat.	lemmas are high
1	A boy is playing soccer.	FA & FC are high
4	The man is reading the newspaper.	lemmas are low
2	A woman is washing clothes.	lemmas are high
8	A person is cutting an apple.	

- NNSO vs NNSLM isn't salient? e.g., NNSO is best for 1, NNSLM for 4; same for 9, 5.
- 5 & 9 involved the most challenging verbs (*raking*, *rowing*); tf-idf approaches work best here, but why?
  - cf. 1 & 4, where frequency measures beat tf-idf