Annotating Picture Description Task Responses for Content Analysis

Anonymous NAACL submission

Abstract

My abstract ...

1 Introduction

Background, purpose of the corpus; (variability within NS and NNS groups; and these groups vary from each other?)

Motivate this for people who don't do this kind of work; measuring 'goodness' of responses; decomposing them into features; (this is not like grammatical error detection or parsing work); Getting a handle on variability, especially under different circumstances; is a response 'native-like'? in what ways (by which features?); breaking variability down into our features;

Cite previous work:
ours (King and Dickinson, 2013)
and others' (Somasundaran et al., 2015)

2 Picture Description Task

The PDT is built around 30 cartoon-like vector graphics. The images were modified to remove any non-essential detail or background. Two images contain numerals, two contain music notes, and one contains a question mark, but the rest are completely devoid of any text or symbols. Each image depicts an ongoing or imminent action, performed by a person or an animal. The images are divided evenly into intransitive, transitive and ditransitive actions.

Two main versions of the PDT were used. In each version, the first half contains "targeted" items, where questions take the form of *What is <subject> doing?*, where the subject (e.g., *the boy*, *the bird*) is provided. The second half contains "untargeted" items, where each question asks *What is happening?*. A roughly equal number of targeted and untargeted responses were collected for each item.

Both halves (targeted and untargeted) are introduced with instructions and an example item and responses. The instructions ask participants to focus on the main event depicted in the image and to respond with a complete sentences. The PDT was presented as an online survey and all participants typed their own responses. Participants were instructed not to use any resources or reference materials, but they were permitted to use browser-based spell checking.

In previous similar work (King and Dickinson, 2013, 2016), NSs were found to produce less variation than NNSs. Many NSs provided the same or very similar response with the most canonical way of expressing the main action. A primary purpose of the current corpus is to assess NNS response content by comparing against the NS responses, so NSs were asked to provide two nonidentical responses in the hopes that this would result in more variation among NS responses.

2.1 Data Collection

PDT versions;
NS vs NNS (1 vs 2 responses);
NNSs demographics (ELIP recruited);
NSs demographics and sources (known persons vs crowdsourced);
Response counts (types, tokens)

3 Annotation

The data were annotated with the aim of providing information that would be useful for the automatic assessment of NNS responses via comparison with the NS responses.

3.1 Scheme

The annotation scheme was developed through an iterative process of annotation, discussion and revision, with input from two annotators and multiple language professionals. The initial scheme

was planned as a three point scale, ranging from *accurate and native-like* to *accurate but not native-like* to *not accurate*. This proved problematic. *Accuracy* and *native-likeness* could not be adequately defined and applied to the data. Eventually, five binary features were settled on, with each feature having some relation to the original concepts of accuracy and native-likeness. A set of annotation guidelines were produced with definitions, rules and examples for each feature. The features and brief descriptions are listed here:

1. **Core Event**: Does the response capture the core event depicted in the image? Core events are not pre-defined but should be fairly obvious given the nature of the images. The response should link an appropriate subject to the event.

2. **Answerhood**: Does the response make a clear attempt to answer the question? This generally requires a progressive verb. For targeted items, the subject of the question must be used as the subject of the response; appropriate pronouns are also acceptable.

3. **Grammaticality**: Is the response free from errors of spelling and grammar?

4. **Interpretability**: Does the response evoke a clear mental image? Any required verb arguments must be present and unambiguous.

5. **Verifiability**: Does the response contain only information that is true and verifiable based on the image? Inferences should not be speculations and are allowed only when necessary and highly probable, as when describing a familial relationship between persons depicted in the image.

3.2 Agreement

Two annotators participated in the annotation. Both are native speakers of English and each has several years of language teaching experience with both children and adult learners. Annotator 1 (A1) annotated the complete corpus. Annotator 2 (A2) annotated only the development set and the test set.

Three items were used as a development set for developing and revising the annotation scheme. These items were also used as examples in the guidelines. They represent one intransitive, one transitive and one ditransitive item. Both annotators annotated the full development multiple times throughout the process, discussing and adjudicating disagreeing annotations before moving on to the test set, which was completed with consultation between the annotatorst.

The test set parallels the development set and consists of one intransitive, one transitive and one ditransitive item, and is shown in Table 1. Agreement and Cohen's kappa scores are given in Table 2.

3.2.1 Intransitives, transitives & ditransitives

Comparisons of the intransitive, transitive and ditransitive items reveal an association between lower item complexity and higher agreement. The highest raw agreement and Cohen's kappa scores are found with the intransitive item, and the lowest are found with the ditransitive item.

This is as expected, as ditransitive sentences are longer with more verb arguments, meaning there are a greater number of opportunities for responses to vary, and thus more opportunities for annotators to disagree on a given response. This trend also matches annotator feedback; both ranked the ditransitive item as the most difficult to annotate (for all features) and the intransitive as the easiest.

LK: A type-token ratio table would be helpful here.

3.2.2 Targeted & untargeted

When the annotations are grouped into targeted and untargeted sets, the raw agreement scores are comparable: 94.9% for targeted and 95.2% for untargeted items. However, despite a greater degree of response variation, the untargeted group has a higher kappa score: 87.2% compared to 82.3%. This may appear counterintuitive, but can most likely be attributed to the stricter annotation rules for targeted items. For example, *answerhood* does not allow targeted responses to modify the subject provided in the question in any way. The untargeted questions elicit a greater degree of variation, but the annotation scheme accommodates this.

When asked to compare the annotation of targeted and untargeted responses, A2 noted that targeted responses require more concentration and closer consultation of the guidelines. Both A1 and A2 describe the annotation of untargeted items as less restrictive.

3.2.3 Features

Grouped by feature, the annotations all show raw agreement scores above 91% and Cohen's kappa

are kappa scores percentages?

scores above 74%. For the future use of this corpus in content assessment, these kappa scores are comfortably above the 67% commonly suggested as a baseline for meaningful, reliable agreement [CITATION].

Core event Likely the most relevant feature here for content assessment, core event has the second lowest kappa score, at 80.8%. This is somewhat lower than expected, as it is lower than the pre-adjudication development set score (88.9%). This appears to be largely attributable to the difficulty of the ditransitive item, which was more challenging for both participants and annotators. The development set item depicts a man delivering a package to a woman, and many participants were familiar with the appropriate vocabulary and constructions. The test set item shows a man giving directions to a woman, and this resulted in a greater degree of variation, as many NNSs did not describe this in a canonical way. Rather than constructions like *asking X for directions* or *giving directions to X*, many NNSs describe the item with phrases like *pointing*, *guiding*, *helping a lost person* or *reading a map*, and most disagreeing core event annotations involve such responses, with A2 more likely to accept these less specific descriptions.

Need to back this up with some numbers!

To a lesser extent, the transitive item, which shows a woman hugging a dog, resulted in disagreements where A2 accepts the word *pet* as the object, but A1 rejects such responses. Despite the reasonable scores for core event agreement, the fact that many disagreements hinge on particular words or annotators having minor differences in interpretation of the event suggest that greater agreement could be easily achieved by providing annotators with suggestions about the acceptable content for each response.

Answerhood This feature was added largely as a way to identify responses that do not follow the instructions. Such responses tend to fall into three categories: responses that simply do not directly answer the given question; responses that are gibberish or very low-effort and entered only so the participant can proceed to the next item; and "troll" responses that attempt to be funny or obscene at the cost of attempting a direct answer.

The majority of participants do attempt to follow the instructions and answer the question, however, and it is unsurprising that this feature skews strongly toward *yes* annotations and results in the highest raw agreement (98.2%) and kappa (93.6%) scores among the five features.

What is the woman doing? [Intrans.]	A1	A2
The woman is running.	1	1
She is wearing a red shirt.	0	0
Trying to run from her bad decisions.	1	0

What is the woman doing? [Transitive]	A1	A2
She's holding a puppy and looks happy.	1	1
She is happy with the dog.	0	0
The woman is loving her dog.	0	1

What is the man doing? [Ditransitive]	A1	A2
giving directions to a woman.	1	1
The man is reading a map.	0	0
The man is is telling her where to go.	1	0

Table 1: Test sample items and example responses with Core Event annotations from Annotators 1 and 2.

Grammaticality

Interpretability

Verifiability Relevant tables:
Types, Tokens, TypeTokenRatio
(for targeted vs. untargeted;

Set	Total	A1Yes	A2Yes	Chance	Agree	Kappa
intrans	2155	0.863	0.855	0.758	0.978	0.910
trans	2155	0.780	0.774	0.653	0.949	0.880
ditrans	2155	0.812	0.786	0.678	0.924	0.764
Target	3390	0.829	0.818	0.709	0.949	0.823
Untarg	3075	0.806	0.790	0.678	0.952	0.872
Core	1293	0.733	0.717	0.601	0.923	0.808
Answer	1293	0.834	0.831	0.721	0.982	0.936
Gramm	1293	0.861	0.872	0.768	0.960	0.827
Interp	1293	0.818	0.787	0.682	0.919	0.744
Verif	1293	0.845	0.817	0.719	0.968	0.884

Table 2: Agreement scores for different groupings of the test set, showing total annotations, *yes* annotations for Annotator 1 and 2, total expected chance agreement (*yes* + *no*), actual raw agreement and Cohen's kappa.

NS response 1 vs. NS response 2)

4 Discussion

4.1 Annotator feedback

Annotators' impressions of the task;
what is difficult? what is easy?

4.2 Agreement & Disagreement Trends

(Do trends align with annotator feedback?)
What kinds of items & responses were challenging?
Are the trends for individual features? Recurring disagreements?

4.3 Limitations

what we'd do differently:
which images are problematic and why (symbols, ambiguity);
Which features are problematic; useful/not useful;

4.4 Potential Uses

Use of corpus in the next phase of my work;
Suggestions of other projects and research questions for this corpus.

Acknowledgments

(Advisors, annotators)

References

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. 2016. Shallow semantic reasoning from an incomplete gold standard for learner language. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 112–121.

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, Denver, Colorado. Association for Computational Linguistics.