Semantic Analysis of Image Based Learner Sentences: Notes regarding statistical analysis

Levi King

October 5, 2017

1 Background and overview

The following notes pertain to the statistical analysis necessary for my dissertation work. The work consists of comparing native speaker (NS) and non-native speaker (NNS) typed responses to a picture description task (PDT). The responses are manually annotated for five binary features. They will also be given a holistic rating (but that rating scale has not yet been determined). The holistic ratings will be used to rank responses. The rankings will be used to gain insights into the binary features, such as which features are most indicative of high or low rankings, and how this differs between NSs and NNSs. Additionally, I will attempt to automatically rate and rank the NNS responses (without the annotation) by using a system that primarily analyzes the frequencies of word n-grams and syntactic dependencies in the NNS and NS responses. A comparison of the automatic rankings and the (manual) holistic rankings will allow for evaluation of the automated approach and should yield more general insights into automatic content assessment for short responses.

2 Data

The dataset consists of responses to PDT items. Each item contains a single, simple cartoon image portraying some common action. 30 images are used, but presented in a "targeted" version, where the participant is asked "What is <the subject> doing?", and in an "untargeted" version, in which the participant is asked, "What is happening?". The images were chosen to represent intransitive, transitive and ditransitive actions equally, with 10 images intended for each of these categories.

NNS participants were recruited from the English Language Improvement Program at Indiana University. NS participants fall into two categories: some were friends or colleagues recruited directly ("Familiar NSs"); a larger number participated in the task through a survey website, where they are rewarded with gift cards or other compensation ("Crowdsourced NSs"). Variation in response quality is readily noticeable between the two NNS groups, and annotation will be used to examine these differences as well. Approximate response counts for each group and setting are presented in Table 1.

		Targeted	Untargeted	Total
NNSs	Tokens	70	70	140
	Types	40	55	95
Familiar NSs	Tokens	40	40	80
	Types	20	30	50
Crowd Src NSs	Tokens	120	120	240
	Types	70	90	160

Table 1: Approximate response counts for a typical item in the set of 30 PDT items.

In a related pilot study, I found that NSs tend to describe common actions with far less variation than NNSs. Because NS responses will be used to form a gold standard (that is, roughly, a model or "answer key") for rating NNS responses, NS participants were required to provide two different responses to each task item. This is intended to elicit a wider range of responses which can be used to match more content in the NNS responses. Thus the NS responses can also be sorted into first responses and second responses, but this distinction is not made in Table 1. A variety of comparisons of first and second NS responses and the quality of automatic NNS response ratings derived from these first and second NS responses may lead to insights regarding the building of a gold standard for similar approaches to automatic content analysis.

3 Feature annotation

Responses are annotated for five binary features. The features have been given names, but these names are used as a shorthand and should not be taken as perfect descriptors of the features, which are laid out in detail in a set of annotation guidelines. A brief description of each is given here.

1. Grammaticality: Without regard to the item context, is the response interpretable as a proposition with a reasonable meaning and *free of grammar or spelling mistakes*?

2. Interpretability: Without regard to the item context, does the response present a proposition that is *described adequately enough to evoke a clear visual image*, without the need to infer important information.

3. Verifiability: Does the response present *only facts that are verifiable by looking at the image*?

4. Core Event: Does the response *describe the main action* taking place in the image?

5. Answerhood: Does the response make an *attempt to answer the question* presented in the item?

As of late September, 2017, Annotator 1 (me) has completed a first pass of roughly two thirds of the annotation for the full set of responses. A second pass will be necessary to make final judgments on responses that were marked “maybe”.

I am training a second annotator. Because this annotation takes considerable effort, my current plan is to have Annotator 2 annotate only a sample of the data and use these annotations to calculate inter rater agreement and gain insights into the annotation scheme itself. This is discussed further below.

4 Holistic ratings

Initially, I planned to compare manual response rankings with automatic response rankings as a way of evaluating automatic approaches to response scoring. The necessary manual annotation was expected to be some single measure on a scale from “accurate and native-like” to “accurate but not native-like” to “not accurate or interpretable”. However, this proved to be overly simplistic, with responses not easily falling on such a continuum, and the annotation scheme expanded to an eventual five binary features. These features almost certainly do not contribute equally to a response’s “goodness”, so there is no obvious way to derive an overall response score from them. Moreover, my automatic response rating system will not provide these binary feature annotations, but will provide a single response score. This means that in order to have some way of evaluating both automatic approaches to rating responses and the importance of each binary feature in response quality, some other holistic annotation is needed. I plan to begin development of a holistic rating scale and accompanying guidelines very soon, as I complete the feature annotation.

I could use advice for establishing a holistic scheme that would be optimally useful for these needs. It seems particularly important to consider the kinds of analysis that would be used when comparing the rankings I will be working with – a "gold standard" ranking based on the manual, holistic scores, and a "test ranking" based on the automatic scores. Each ranking will consist of between 100 and 300 unique responses. The test ranking will be based on the automatic scores given to each response; in the current implementation, this score is some value between 0 and 1. Please note also that there will in fact be multiple gold standard rankings, based on responses from different sets of participants, e.g., Familar NNSs vs Crowd Sourced NNSs, as well as multiple test rankings, based on varying the parameters in the automatic approach.

In informal discussions with colleagues, I've been told that I'll likely want to use some form of logistic regression (or probit regression?) for comparing continuous values (automatic response scores) with categorical values (the manual holistic scores). I assume that the holistic scheme should ideally use the minimum number of levels necessary for capturing the desired distinctions. Are there other considerations I should take before designing this holistic scheme?

5 Inter rater reliability (or similar measures)

I plan to have Annotator 2 annotate three items for all binary features and for the holistic score. These items would consist of one intransitive, one transitive and one ditransitive action, and would represent 10% of the responses. However, I would especially appreciate suggestions with regard to the appropriate amount of annotation needed from Annotator 2 in order to yield meaningful analysis. I know that a number of similar analyses could be considered for this scenario. In particular, I could use suggestions on the most appropriate analysis and how it might best be implemented and interpreted (i.e., Excel, SPSS, etc.).