Institute of Informatics
Logics and Security Studies

Analysis of Annotator Agreement in Online Chat Discussions

Patrick Zongo, Samira Shaikh
09/18/2009

ILS Tech Report: 005

Available for download from:
`http://www.ils.albany.edu`

Abstract

Developing efficient data mining systems in data mining is contingent on many specific characteristics. One of the most important aspects is being able to develop reliable training data for software to learn on. For our current projects we have been creating annotated chat room data. The following paper attempts to briefly describe the characteristics of the annotated data and the reliability of the annotators.

1 Data Collection

In order to collect relevant data we have obtained 14 sessions of directed chat room text. Within these chat sessions the amount of participants has ranged from three to nine participants, each of whom contributed to approximately 400 utterances per chat session.

After collecting this data we selected annotators to analyze each utterance to determine notable characteristics of the data. These characteristics included the topic that the utterance was about, the focus of the utterance within that topic, the dialog act, the utterance that was being referred to, et al.

2 Annotator Agreement

To develop a baseline for our annotator analysis we first created a group of internal annotators who were familiar with the characteristics of annotation that we sought. As a means of verifying how reliable all other annotators were the internal annotators annotated the March 22nd, 2009 chat session, which contained 468 utterances.

2.1 Dialog Act Analysis

Within this chat session many characteristics were annotated, but for our analysis we focused on the annotation of dialog acts. The reason for this selection is that the set of dialog acts is finite, thus it can serve as a more well-defined measure of how much the internal annotators agree.

To determine their interannotator agreement we used the following metrics:

Complete Agreement is the percentage of utterances on which there were was complete agreement between the annotators. Complete agreement is achieved when all of the internal annotators and the external annotator agreed on a specific utterance. (Range: 0 to 100%)

Partial Agreement is the percentage of utterances on which a simple majority(more than 50%) of the annotators agreed. Both the Partial Agreement and the Complete Agreement is dependent on the amount of internal annotators who agreed on the topic. (Range: 0 to 100%)

Fleiss' Kappa [Fle71] is a measurement of annotator reliability above that expected by chance. (Range: -1 to 1)

Krippendorff's Alpha [Kri80] is similar to Fleiss' Kappa except that it accounts for incomplete data. (Range: -1 to 1)

Conflated Krippendorff's Alpha is an extension of Krippendorff's Alpha. A benefit of Krippendorff's Alpha is the ability to modify by how much annotators' annotations differ. In this Conflated Krippendorff's Alpha we use a mapping from some of the tags to a parent tag(See Table 1). This is used when an ambiguous situation could warrant the use of more than one tag.(Range: -1 to 1)

Ground Agreement is the percentage of agreement with the "ground truth". To compute the ground agreement we first iterated through the internal annotator's annotations to develop a ground truth. This ground truth is determined by the following process: when more than half of the internal annotators agree on a specific annotator it gets deemed as the ground truth.(Range: 0 to 100%)

Conflated Ground Agreement is similar to the Ground Agreement with the added benefit of using the mapping of dialog act tags that was used for the Conflated Krippendorff's Alpha.(Range: 0 to 100%)

In addition to internal annotators we have also chosen external annotators who periodically annotate different chat room sessions. So far, six annotators have annotated sessions, but more annotations should be forthcoming.

The quality of external annotators is important for developing a test set of our annotations. To determine the quality of external annotators they are all expected to annotate the March 22nd chat session. This gives us

Table 1: Mapping of Conflated Tags

Parent Tag	Child Tag	Child Tag	Child Tag
ASSERTION	assertion-opinion	explanation	
DISAGREE-REJECT	disagree-reject	maybe-hold	
AGREE-ACCEPT	agree-accept	maybe-hold	
ACCEPT-PART	accept-part	reject-part	
ACKNOWLEDGE	acknowledge	acknowledge-correct	backchannel
SIGNAL-NON-UNDERSTANDING	signal-non-understanding	repeat-rephrase	
COMMUNICATION-MANAGEMENT	communication-management	conventional-response	

the opportunity to fairly assess their annotation quality against the internal annotators, who have all annotated this chat session. [1]

2.2 Topic Analysis

Another component of dialog act annotation that we would like to investigate is the nature of topic extraction in the chat sessions. The annotators that we have chosen are not given a specific subset of topics or foci within which to classify an utterance. This creates an ample amount of variability in the specific words used to describe a topic. For this reason, we will only be concerned with whether or not each utterance is classified as a new topic.

Withing each topic annotated, we have also allowed an extra degree of freedom for the annotators to select a focus within that topic. This creates the opportunity to better specify when some form of change has occurred in the chat session. A simplified example can be seen in Table 4. In this segment of chat, the annotator noticed that while the conversation's topic is travel, Amanda momentarily switches the focus to her own travel preferences as opposed to Japan. Similar to topics we will be focusing on whether or not the focus is new and not on the designated name of each focus.

[1] It should be noted that in Table 2 the complete and partial agreement fields are very sensitive to the amount of annotators being compared.

Table 2: Agreement between internal annotators, external annotators, and individual external annotators alongside internal annotators

Metric	Caesar	Farina	Jingsi	Kate	Michelle	Rhiannon
observed	52.30%	52.65%	48.79%	49.25%	42.52%	47.45%
chance	19.52%	19.75%	17.13%	18.05%	16.03%	19.07%
Complete Agreement	26.34%	25.54%	20.77%	23.66%	14.99%	23.13%
Fleiss' Kappa	40.74%	40.99%	38.20%	38.07%	31.55%	35.06%
Krippendorff's Alpha	41.13%	41.34%	38.60%	38.55%	31.80%	35.46%
Ground Agreement	55.89%	56.22%	46.04%	48.82%	31.48%	39.61%
Conflated Ground Agreement	59.10%	64.38%	51.18%	60.86%	50.11%	55.89%
Conflated Krippen-dorff's Alpha	45.92%	48.15%	43.85%	46.77%	42.59%	44.17%
Partial Agreement	65.74%	66.52%	61.24%	61.08%	52.68%	52.68%

Table 3: Agreement between internal annotators and external annotators using various metrics

Metric	Internal Annotator Agreement	External Annotator Agreement
observed	52.78%	44.49%
chance	21.85%	11.28%
Fleiss' Kappa	39.59%	37.43%
Complete Agreement	36.19%	10.99%
Partial Agreement	86.51%	55.60%
Krippendorff's Alpha	40.10%	37.61%
Conflated Krippen-dorff's Alpha	45.43%	43.23%
Amount of Annotators	3	6

Table 4: Utterances 57 to 63 of the March 22nd 2009 chat session

Speaker	Focus	Topic	Utterance
Ken	japan	travel	i lived in japan for 7 years
Kara	japan	travel	cool!
Nick	japan	travel	what part?
Kara	japan	travel	where in Japan?
Amanda	travel pref-erences	travel	Well, I'mo not in a traveling mood. I laways want to sleep in my own bed.
Ken	japan	travel	me?
Nick	japan	travel	yeah, where in Japan?
Ken	japan	travel	ah, kyoto for 1 year, tokyo for 6

Table 5: Krippendorff's Alpha measurement for "New Topic" or "New Focus" and just "New Topic"

Annotator	New Topic or Focus	New Topic
Caesar	40.07%	26.68%
Farina	48.60%	44.99%
Jingsi	43.48%	40.92%
Kate	incomplete	incomplete
Michelle	incomplete	incomplete
Rhiannon	34.68%	39.44%

Due to the training the annotators were given, they have an amount of leeway which could cause some difficulties in determining whether or not the transition in the conversation flow is a result of a change in topic or a change in focus. For this reason, I have done a comparison of two distinct situations: whether or not only the topic is new; and whether or not either the topic or the focus is new. The results of this analysis can be found in Table 5. The annotators in question were judged alongside of the internal annotators. Their agreement was determined by measuring Krippendorff's Alpha.

3 Future Analysis

As new annotators enter this project we plan to have them analyze the same chat session as a baseline. With these new annotators we can better understand the nature of the annotators and the quality of their assessments. Due to the potentially high Krippendorff's Alpha for topic transition, in the future we will analyze topic transitions and attempt to predict similarity of topic names that are chosen by different annotators.

References

[Fle71] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 74(5):378–382, 1971.

[Kri80] Klaus Krippendorff. *Content Analysis, an Introduction to its Methodology.* Sage Publications, Thousand Oaks, CA, 1980.