We thank the anonymous reviewers for the valuable comments and suggestions. Below are our responses to some of the questions:

Reviewer 1:

1. “k arguments per each verb approach unconvincing”: We gave a detailed discussion on this issue in the last paragraph of section 3. In short, the granularity of senses for any given verb is not fixed, which is evident from different English dictionaries. Therefore, we decide to enable the system to control the granularity. For simplicity of evaluation, we set k = 5, 10, or 15 for 50 verbs, but in practice, the number k can be set differently for different verbs, which we view as an advantage of the framework.
2. “deriving k from the argument count from FrameNet or Propbank”: This is often not reasonable, because FrameNet or Propbank does not necessarily have the right granularity of senses for a verb. Taking “enjoy” as an example, FrameNet only contains one concept “Stimulus” as object concept. But if we set k as 5, we may produce a more precise set of concepts. Some examples are in Table 2 (page 7).
3. “<verb, subject> should be added”: We have conducted experiments on subjects, too. However, since the verb-object relation better reflects the semantics of the verbs, and due to space limitation, we did not show these results in the paper, but included them in the project website (footnote of page 5). We will include them in the revised version.
4. “evaluation does not consider any hypernym or synonym”: We do not consider hypernyms because they are often incorrect. For example, in Probase, besides “instrument”, “piano” has hypernym “skill”, “class”, etc., which cannot be “played”. We do consider the synonyms in WordNet, collectively represented by a synset, and all words in the synset are considered correct arguments. Probase does not have synsets, and therefore does not consider synonyms.

Reviewer 2:

1. As discussed in section 4.2, we use entropy to evaluate whether an argument is informative w.r.t. different patterns: higher entropy means the argument appears in more patterns and hence is of better quality. We use mutual information to compute the relatedness between an argument and a verb because the more information they overlap, the more related they are.
2. The statement “an argument that appears in more patterns has higher probability to be correct, and thus has higher quality” is based on statistics. We are not saying that any argument that matches just one pattern must be wrong. Certainly there may be such negative cases. However, according to our observation, correct arguments generally have more patterns than incorrect ones. For example, as the object of “eat”, “corn” has 142 patterns and “habit” has 42 patterns. We postulate that the arguments that appear in only one or a small number of patterns are often idioms or parsing errors.
3. The three human annotators are native English speakers. We also release the annotated data sets in the project website (footnote of page 5).
4. As explained in section 3, L is the set of edges between every two concepts that have overlap less than a threshold \tau (page 2).
5. In sec 5.3 (not 4.3), we rely on the isA relation in the taxonomy (Probase or Wordnet) to determine if an argument belongs to \*any\* of the k argument concepts. If so, that argument is identified to be correct.
6. Regarding input parameter k, please refer to response 1 to Reviewer 1.

Reviewer 3:

1. The unsightly hyphenation is our negligence and will be corrected in the revision.