Embracing Ambiguity:
A Comparison of Annotation Methodologies for Crowdsourcing Word Sense Labels

UCLA
Sapienza Università di Roma

David Jurgens
jurgens@di.uniroma1.it

Introduction and Motivation

Many NLP applications require knowing which sense of a word is present in a context. Gathering sense annotations is time consuming so crowdsourcing is often used. However, sense annotation is often very difficult for untrained annotators due to confusion over which senses apply. This confusion leads to low annotator agreement and lower quality annotations.

Objective

We want to develop annotation methodologies that achieve high agreement when used with untrained annotators in a crowdsourcing setting. Our hypothesis is that higher agreement can be obtained by encouraging annotators to use all the senses they think are applicable, thereby making any ambiguity explicit and measurable.

Methodology

Compare the traditional single-sense annotation method with three multi-sense methods.

Re-annotate the same contexts as Erk et al. (2009), who gathered Likert ratings for each sense of 8 words on 50 contexts each.

Measure annotator agreement for each method using Krippendorff's α, where 1 indicates complete agreement.

Use Amazon Mechanical Turk (MTurk) to gather 10 annotations per instance or 3 per sense combination for MaxDiff.

Likert

Rate each sense by its applicability

The student handed her **paper** to the professor

1	2	3	4	5	
○	○	◉	○	○	A material made of cellulose pulp
○	○	○	○	◉	An essay
◉	○	○	○	○	A daily or weekly publication
◉	○	○	○	○	A medium for written communication
◉	○	○	○	○	A scholarly article
◉	○	○	○	○	A business firm that publishes newspapers

- Originally used by Erk et al. (2009)
- Constant number of MTurk tasks per context
- Straight-forward to use

Select and Rate

Select which senses might apply

The student handed her **paper** to the professor

- ☑ A material made of cellulose pulp
- ☑ An essay
- ☐ A daily or weekly publication
- ☐ A medium for written communication
- ☐ A scholarly article
- ☐ A business firm that publishes newspapers

Rate selected senses by their applicability

The student handed her **paper** to the professor

1	2	3	4	5	
○	○	◉	○	○	A material made of cellulose pulp
○	○	○	○	◉	An essay

- Only **Rate** senses that pass a **Select** threshold
- Easier to annotate very polysemous words

MaxDiff

Select the senses whose meanings *most* and *least* apply

The student handed her **paper** to the professor

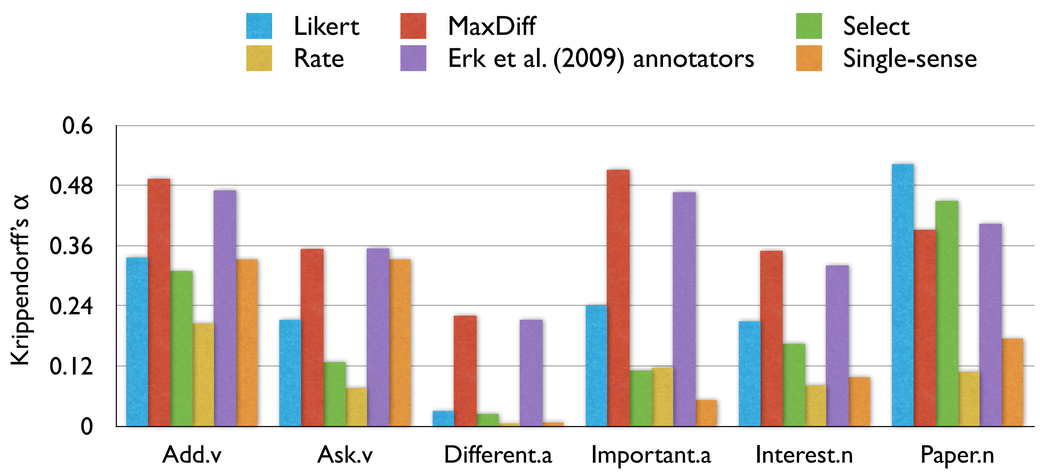
Most		*Least*
◉	A material made of cellulose pulp	○
○	A scholarly article	○
○	A business firm that publishes newspapers	◉

- Removes rating scale bias
- Converted into numeric ratings by aggregating
- Number of MTurk tasks scales with the number of senses

Experiment 1

Do annotators agree with each other?

Methodology:
Measure Krippendorff's α for each word's annotations from each annotation method

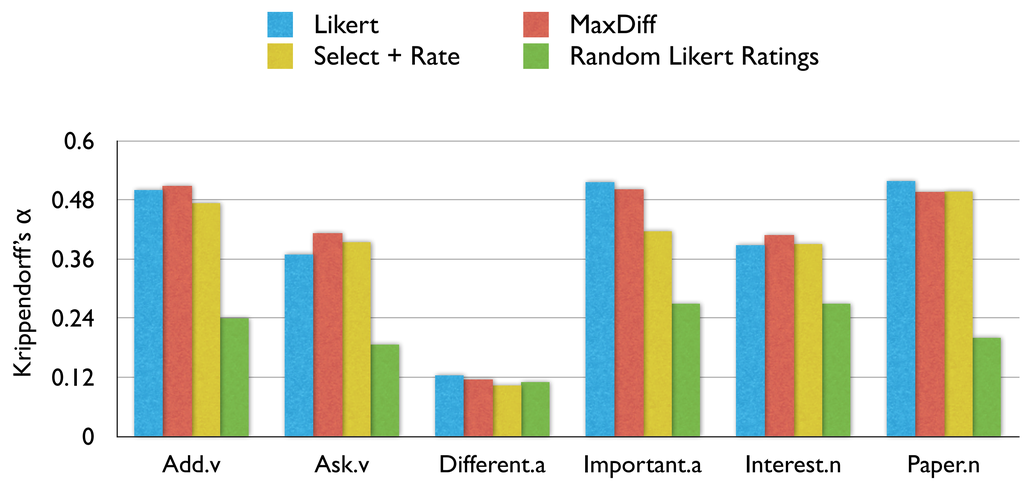


Result: Annotating with multiple sense provides a substantial improvement in agreement over using a single sense. MaxDiff performs best on average

Experiment 2

Can we improve agreement by aggregating ratings?

Methodology:
Compute an average sense rating by combining the MTurk annotations. Then measure Krippendorff's α with Erk et al.'s annotators

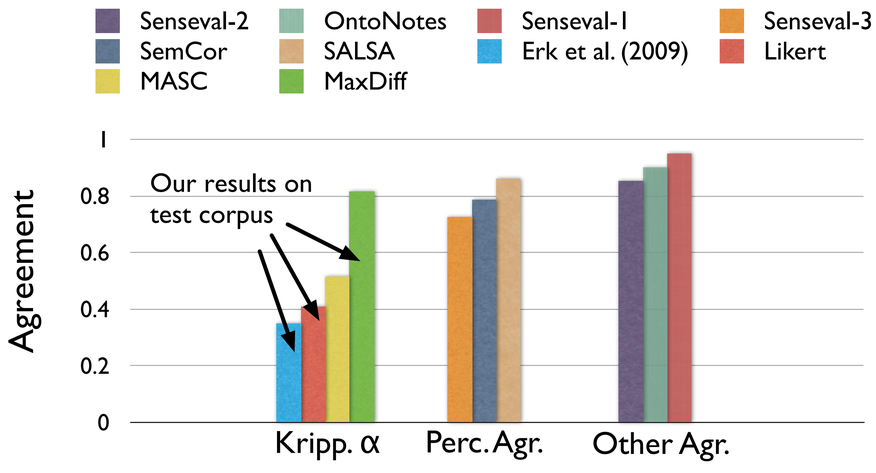


Result: Combining sense ratings results in substantial increases in agreement. Likert and MaxDiff methods perform well.

Experiment 3

How reproducible are the annotations?

Methodology:
Sample two independent sets of ratings for each context. Then measure Krippendorff's α between the combined ratings from each sample.



Result: Independent aggregated MaxDiff sense annotations have an agreement level on par with expert annotators in existing corpora

Conclusions

Allowing annotators to use multiple senses generates annotations with much higher agreement than if annotators were restricted to using a single sense.

Aggregating multi-sense annotations into one sense rating improves quality.

MaxDiff is highly replicable and has agreement consistent with that from high-quality sense annotations by expert lexicographers.

Amazon Mechanical Turk is a viable platform for gathering sense annotations when using a fine-grained sense inventory such as WordNet.

References

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In Proceedings of LREC.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of ACL*, pages 10–18. ACL.

Christiane Fellbaum, Jaochim Grabowski, and Shari Landes. 1998. Performance and confidence in a semantic annotation task. WordNet: An electronic lexical database, pages 217–237.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of NAACL*, pages 57–60. ACL.

Adam Kilgarriff. 1999. 95% replicability for manual word sense tagging. In Proceedings of EACL, pages 277–278. ACL.

Adam Kilgarriff. 2002. English lexical sample task description. In Senseval-2: Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems.

Rebecca J. Passonneau, Ansaf Salleb-Aoussi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In Proceedings of LREC.