## 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.

## 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
- 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)
- 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

- 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?**

- Likert
- Rate
- MaxDiff
- Erk et al. (2009) annotators
- Select
- Single-sense

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

![Results](image)

**Result:** Announcing with multiple senses 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?**

- Likert
- Select + Rate
- MaxDiff
- Random Likert Ratings

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

![Results](image)

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

---

### Experiment 3

**How reproducible are the annotations?**

- Senseval-2
- OntoNotes
- MASCC
- Senseval-1
- Senseval-3
- Erk et al. (2009)
- Likert

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

![Results](image)

**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