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

David Jurgens
Department of Computer Science
University of California, Los Angeles
jurgens@cs.ucla.edu

Abstract

Word sense disambiguation aims to identify which meaning of a word is present in a given usage. Gathering word sense annotations is a laborious and difficult task. Several methods have been proposed to gather sense annotations using large numbers of untrained annotators, with mixed results. We propose three new annotation methodologies for gathering word senses where untrained annotators are allowed to use multiple labels and weight the senses. Our findings show that given the appropriate annotation task, untrained workers can obtain at least as high agreement as annotators in a controlled setting, and in aggregate generate equally as good of a sense labeling.

1 Introduction

Word sense annotation is regarded as one of the most difficult annotation tasks (Artstein and Poesio, 2008) and building manually-annotated corpora with high-quality sense labels is often a time- and resource-consuming task. As a result, nearly all sense-tagged corpora in wide-spread use are created using trained annotators (Hovy et al., 2006; Passonneau et al., 2010), which results in a knowledge acquisition bottleneck for training systems that require sense labels (Gale et al., 1992). In other NLP areas, this bottleneck has been addressed through gathering annotations using many untrained workers on platforms such as Amazon Mechanical Turk (MTurk), a task commonly referred to as crowdsourcing. Recently, several works have proposed gathering sense annotations using crowdsourcing (Snow et al., 2008; Biemann and Nygaard, 2010; Passonneau et al., 2012b; Rumshisky et al., 2012). However, these methods produce sense labels that are different from the commonly used sense inventories such as WordNet (Fellbaum, 1998) or OntoNotes (Hovy et al., 2006). Furthermore, while Passonneau et al. (2012b) did use WordNet sense labels, they found the quality was well below that of trained experts.

We revisit the task of crowdsourcing word sense annotations, focusing on two key aspects: (1) the annotation methodology itself, and (2) the restriction to single sense assignment. First, the choice in sense inventory plays an important role in gathering high-quality annotations; fine-grained inventories such as WordNet often contain several related senses for polysemous words, which untrained annotators find difficult to correctly apply in a given context (Chugur et al., 2002; McCarthy, 2006; Palmer et al., 2007; Rumshisky and Batiukova, 2008; Brown et al., 2010). However, many agreement studies have restricted annotators to using a single sense, which can significantly lower inter-annotator agreement (IAA) in the presence of ambiguous or polysemous usages; indeed, multiple studies have shown that when allowed, annotators readily assign multiple senses to a single usage (Véronis, 1998; Murray and Green, 2004; Erk et al., 2009; Passonneau et al., 2012b). Therefore, we focus on annotation methodologies that enable workers to use as many labels as they feel appropriate, asking the question: if allowed to make labeling ambiguity explicit, will annotators agree? Furthermore, we adopt the goal of Erk et al. (2009), which enabled annotators to weight each sense by its applicability to the given context, thereby quantifying the ambiguity.

Proceedings of NAACL-HLT 2013, pages 556–562,
Atlanta, Georgia, 9–14 June 2013.

This paper provides the following contributions. First, we demonstrate that the choice in annotation setup can significantly improve IAA and that the labels of untrained workers follow consistent patterns that enable creating high quality labeling from their aggregate. Second, we find that the sense labeling from crowdsourcing matches performance with annotators in a controlled setting.

2 Related Work

Given the potential utility of a sense-labeled corpus, multiple studies have examined how to efficiently gather high quality sense annotations. Snow et al. (2008) had MTurk workers, referred to as Turkers, disambiguate uses of "president." While they reported extremely high IAA (0.952), their analysis was only performed on a single word.

Biemann and Nygaard (2010) and Biemann (2012) construct a sense-labeled corpus by concurrently constructing the sense inventory itself. Turkers used a lexical substitution task to identify valid substitutions of a target word. The contexts for the resulting substitutions were clustered based on their word overlap and the resulting clusters were labeled as senses. Biemann and Nygaard (2010) showed that the number of sense definitions for a word in their inventory was correlated with the number in WordNet, often with their inventory having fewer senses by combining related meanings and omitting rare meanings.

Hong and Baker (2011) evaluated multiple annotation strategies for gathering FrameNet sense annotations, ultimately yielding high (>90%) accuracy for most terms after filtering. They highlight ambiguous and polysemous usages as a notable source of errors, which the present work directly addresses.

In the most related work, Passonneau et al. (2012b) had Turkers annotate contexts using one or more senses, with the requirement that a worker labels all contexts. While they found that agreement between all workers was low, their annotations could be combined using the GLAD model (Whitehill et al., 2000) to obtain good performance, though not as good as trained annotators.

3 Annotation Methodologies

We consider three methodologies for gathering sense labels: (1) the methodology of Erk et al. (2009) for gathering weighted labels, (2) a multi-stage strategy that uses both binary and Likert ratings, and (3) MaxDiff, a paired choice format.

Likert Ratings Likert rating scales provide the most direct way of gathering weighted sense labels; Turkers are presented with all senses of a word and then asked to rate each on a numeric scale. We adopt the annotation guidelines of Erk et al. (2009) which used a five-point scale, ranging from 1 to 5, indicating the sense does not apply or that it matches the contextual usage exactly, respectively.

Select and Rate Recent efforts in crowdsourcing have proposed multi-stage processes for accomplishing complex tasks, where efforts by one group of workers are used to create new subtasks for other workers to complete (Bernstein et al., 2010; Kittur et al., 2011; Kulkarni et al., 2012). We propose a two-stage strategy that aims to reduce the complexity of the annotation task, referred to as Select and Rate (S+R). First, Turkers are presented with all the senses and asked to make a binary choice of which senses apply. Second, a Likert rating task is created for only those senses whose selection frequency is above a threshold, thereby concentrating worker focus on a potentially smaller set of senses.

Our motivation for S+R is two-fold. First, the sense definitions of certain words may be unclear or misinterpreted by a minority of the Turkers, who then systematically rate inapplicable senses as applicable. The Select task can potentially remove such noise and therefore improve both IAA and rating quality in the subsequent Rate task. Second, while the present study analyzes words with 4–8 senses, we are ultimately interested in annotating highly polysemous words with tens of senses, which could present a significant cognitive burden for an annotator to rate concurrently. Here, the Select stage can potentially reduce the number of senses presented, leading to less cognitive burden in the Rate stage. Furthermore, as a pragmatic benefit, removing inapplicable senses reduces the visual space required for displaying the questions on the MTurk platform, which can improve annotation throughput.

MaxDiff MaxDiff is an alternative to scale-based ratings in which Turkers are presented with a only subset of all of a word's senses and then asked to select (1) the sense option that *best* matches the mean-

	add.v	ask.v	win.v	argument.n	interest.n	paper.n	different.a	important.a
Erk et al. (2009) IAA	0.470	0.354	0.072	0.497	0.320	0.403	0.212	0.466
MTurk Likert IAA	0.336	0.212	0.129	0.250	0.209	0.522	0.030	0.240
MTurk Select	0.309	0.127	0.179	0.192	0.164	0.449	0.024	0.111
MTurk Rate	0.204	0.076	0.026	0.005	0.081	0.108	0.005	0.116
MTurk MaxDiff	0.493	0.353	0.295	-	0.349	0.391	0.220	0.511
Likert Mode	0.500	0.369	0.083	0.445	0.388	0.518	0.124	0.516
S+R Median	0.473	0.394	0.149	0.497	0.390	0.497	0.103	0.416
MTurk MaxDiff	0.508	0.412	0.184	-	0.408	0.496	0.115	0.501
Sampled Baseline	0.238	0.178	0.042	0.254	0.162	0.205	0.100	0.221
Random Baseline	0.239	0.186	0.045	0.249	0.269	0.200	0.110	0.269

Table 1: IAA per word (top) and IAA between aggregate labelings and the GWS annotators (bottom)

ing in the example context and (2) the sense option that *least* matches (Louviere, 1991). In our setting, we presented three options at a time for words with fewer than seven senses, and four options for those with seven senses. For a single context, multiple subsets of the senses are presented and then their relative ranking is used to produce the numeric rating. The final applicability ratings were produced using a modification of the counting procedure of Orme (2009). First, all sense ratings are computed as the number of times the sense was rated *best* minus the number of times rated *least*. Second, all negatively-rated senses are assigned score of 1, and all positively ratings are normalized to be $(1, 5]$.

4 Experiments

For measuring the difference in methodologies, we propose three experiments based on different analyses of comparing Turker and non-Turker annotations on the same dataset, the latter of which we refer to as the reference labeling. First, we measure the ability of the Turkers individually by evaluating their IAA with the reference labeling. Second, many studies using crowdsourcing combine the results into a single answer, thereby leveraging the wisdom of the crowds (Surowiecki, 2005) to smooth over inconsistencies in the data. Therefore, in the second experiment, we evaluate different methods of combining Turker responses into a single sense labeling, referred to as an *aggregate* labeling, and comparing that with the reference labeling. Third, we measure the replicability of the Turker annotations (Kilgarriff, 1999) using a sampling methodology. Two equally-sized sets of Turker annotations are created by randomly sampling without replacement from the full set of annotations for each item. IAA is calculated between the aggregate labelings computed from each set. This sampling is repeated 50 times and we report the mean IAA as a measure of the expected degree of replicability when annotating using different groups of Turkers.

For the reference sense labeling, we use a subset of the GWS dataset of Erk et al. (2009), where three annotators rated 50 instances each for eight words. For clarity, we refer to these individuals as the GWS annotators. Given a word usage in a sentence, GWS annotators rated the applicability of all WordNet 3.0 senses using the same Likert scale as described in Section 3. Contexts were drawn evenly from the SemCor (Miller et al., 1993) and SENSEVAL-3 lexical substitution (Mihalcea et al., 2004) corpora. GWS annotators were apt to use multiple senses, with nearly all instances having multiple labels.

For each annotation task, Turkers were presented with an identical set of annotation guidelines, followed by methodology-specific instructions.[1] To increase the familiarity with the task, four instances were shown per task, with all instances using the same target word. Unlike Passonneau et al. (2012b), we did not require a Turker to annotate all contexts for a single word; however many Turkers did complete the majority of instances. Both the Likert, Select, and Rate tasks used ten Turkers each. Senses were passed from Select to Rate if they received at

[1]Full guidelines are available at `http://cs.ucla.edu/~jurgens/sense-annotation/`

least three votes. For MaxDiff, we gathered at least $3n$ annotations per context where n is the number of senses of the target word, ensuring that each sense appeared at least once. Due to resource limitations, we omitted the evaluation of *argument.n* for MaxDiff. Following the recommendation of Kosinski et al. (2012), Turkers were paid \$0.05USD for each Likert, Select, and Rate task. For MaxDiff, due to their shorter nature and comparably high volume, Turkers were paid \$0.03USD per task.

To ensure fluency in English as well as reduce the potential for low-quality results, we prefaced each task with a simple test question that asked the Turker to pick out a definition of the target word from a list of four options. The incorrect options were selected so that they would be nonsensical for anyone familiar with the target word. Additionally, we rejected all Turker responses where more than one option was missing a rating. In the case of missing ratings, we infer a rating of 1. Approximately 20-30% of the submissions were rejected by these criteria, underscoring the importance of filtering.

For measuring IAA, we selected Krippendorff's α (Krippendorff, 1980; Artstein and Poesio, 2008), which is an agreement coefficient that handles missing data, as well as different levels of measurement, e.g., nominal data (Select and MaxDiff) and interval data (Likert and Rate).[2] Krippendorff's α adjusts for chance, ranging between $[-1, 1]$ for nominal data and $(-1, 1]$ for interval data, where 1 indicates perfect agreement and -1 indicates systematic disagreement; random labels would have an expected α of zero. We treat each sense and instance combination as a separate item to rate.

5 Results

The results of the first experiment appear in the top of Table 1. Two important aspects emerge. First, the word itself plays a significant role in IAA. Though Erk et al. (2009) reported a pair-wise IAA of the GWS annotators between 0.466 and 0.506 using Spearman's ρ, the IAA varies considerably between words for both Turkers and GWS annotators when measured using Krippendorff's α.

Second, the choice of annotation methodology significantly impacts IAA. While both the Likert and S+R tasks have lower IAA than the GWS annotators do, the MaxDiff annotators achieve higher IAA for almost all words. We hypothesize that comparing senses for applicability is an easier task for the untrained worker, rather than having to construct a mental scale of what constitutes the applicability of each sense. Surprisingly, the binary Select task has a lower IAA than the more complex the Likert task. An analysis of the duration of median task completion times for the Likert and Select tasks showed little difference (with the exception of *paper.n*, which was on average 50 second faster for Likert ratings), suggesting that both tasks are equally as cognitively demanding. In addition, the Rate task has the lowest IAA, despite its similarity to the Likert task. An inspection of the annotations shows that the full rating scale was used, so the low value is not due to Turkers always using the same rating, which would yield an IAA near chance.

In the second experiment, we created a aggregate sense labeling and compared its IAA with the GWS annotators, shown in Table 1 (bottom). For scale-based ratings, we considered three arithmetic operations for selecting the final rating: mode, median, and mean. We found that the mode yielded the highest average IAA for the Likert ratings and median for S+R; however, the differences in IAA using each operation were often small. We compare the IAA with GWS annotators against two baselines: one generated by sampling from the GWS annotators' rating distribution, and a second generated by uniformly sampling in $[1, 5]$. By comparison, the aggregate labelings have a much larger IAA than the baselines, which is often at least as high as the IAA amongst the GWS annotators themselves, indicating that the Turkers in aggregate are capable of producing equivalent ratings. Of the three annotation methodologies, MaxDiff provides the highest IAA both within its annotators and with its aggregate key. Surprisingly, neither the Likert or S+R aggregate labeling appears better than the other.

Based on the second experiment, we measured the average IAA across all words between the aggregate Likert and MaxDiff solutions, which was 0.472. However, this IAA is significantly affected by the annotations for *win.v* and *different.a*, which had the lowest IAA among Turkers (Table 1) and there-

[2]We note that although the ratings are technically given on an ordinal scale (ranks), we use the interval scale to allow comparison with rational ratings from the aggregate solutions.

Corpus	Sense Inventory	IAA	Measurement
SensEval-1 (Kilgarriff and Rosenzweig, 2000)	HECTOR	0.950	Replicability experiment (Kilgarriff, 1999)
OntoNotes (Hovy et al., 2006)	OntoNotes	$\geq 0.90^{\dagger}$	Pair-wise agreement
SALSA (Burchardt et al., 2006)	FrameNet	0.86	Percentage agreement
SensEval-2 Lexical Sample (Kilgarriff, 2002)	WordNet 1.7	0.853, 0.710, 0.673‡	Adjudicated Agreement
GWS with MaxDiff Replicability$^{\diamond}$	WordNet 3.0	0.815	Krippendorff's α
SemCor (Fellbaum et al., 1998)	WordNet 1.6	0.786, 0.57*	Percentage agreement
SensEval-3 (Snyder and Palmer, 2004)	WordNet 1.7	0.725	Percentage agreement
MASC (Passonneau et al., 2012a)	WordNet 3.1	-0.02 to 0.88$^{\triangleleft}$	Krippendorff's α with MASI (Passonneau et al., 2006)
MASC, single phase reported in Passonneau et al. (2010)	WordNet 3.1	0.515	Krippendorff's α
GWS with Likert Replicability	WordNet 3.0	0.409	Krippendorff's α
GWS with Erk et al. (2009) annotators	WordNet 3.0	0.349	Krippendorff's α

† Not all words achieved this agreement.

‡ Kilgarriff (2002) uses a multi-stage agreement procedure where two annotators rate each item, and in the case of disagreement, a third annotator is added. If the third annotator agrees with either of the first two, the instance is marked as a case of agreement. However, the unadjudicated agreement for the dataset was 67.3 measured using pair-wise agreement. A re-annotation by Palmer et al. (2004) produced a similar pair-wise agreement of 71.0.

* Tou et al. (1999) perform a re-annotation test of the same data using student annotators, finding substantially lower agreement

$^{\diamond}$ Excludes agreement for *argument.n*, which was not annotated

$^{\triangleleft}$ IAA ranges for 37 words; no corpus-wide IAA is provided.

Table 2: IAA for sense-annotated corpora

fore produce noisy aggregate solutions. When *win.v* and *different.a* are excluded, the agreement between aggregate Likert and MaxDiff solutions is 0.649. While this IAA is still moderate, it suggests that Turkers can still produce similar annotations even when using different annotation methodologies.

For the third experiment, replicability is reported as the average IAA between the sampled aggregate labelings for all annotated words. Table 2 shows this IAA for Likert and MaxDiff methodologies in comparison to other sense annotation studies. Krippendorff (2004) recommends that an α of 0.8 is necessary to claim high-quality agreement, which is achieved by the MaxDiff methodology. In contrast, the average IAA between sampled Likert ratings is significantly lower, though the methodology does achieve an α of 0.812 for *paper.n*. However, when the two words with the lowest IAA, *win.v* and *different.a*, are excluded, the average α increases to 0.880 for MaxDiff and 0.649 for Likert. Overall, these results suggest that MaxDiff can generate highly replicable annotations with agreement on par with that of other high-quality sense-labeled corpora. Furthermore, the Likert methodology may in aggregate still produce moderately replicable annotations in some cases.

6 Conclusion and Future Work

Word sense disambiguation is a difficult task, both for humans and algorithms, with an important bottleneck in acquiring large sense annotated corpora. As a potential solution, we proposed three annotation methodologies for crowdsourcing sense labels. Importantly, we relax the single sense assignment restriction in order to let annotators explicitly note ambiguity through weighted sense ratings. Our findings reveal that moderate IAA can be obtained using MaxDiff ratings, with IAA surpassing that of annotators in a controlled setting. Furthermore, our findings showed marked differences in rating difficulty per word, even in the weighted rating setting. In future work, we will investigate what factors influence annotation difficulty in order to improve IAA to what is considered expert levels, drawing from existing work analyzing difficulty in the single label setting (Murray and Green, 2004; Passonneau et al., 2009; Cinková et al., 2012).

References

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Michael S. Bernstein, Ggreg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of UIST*, pages 313–322. ACM.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing WordNet. In *The 5th International Conference of the Global WordNet Association (GWC-2010)*.

Chris Biemann. 2012. Turk Bootstrap Word Sense Inventory 2.0: A Large-Scale Resource for Lexical Substitution. In *Proceedings of LREC*.

Susan Windisch Brown, Travis Rood, and Martha Palmer. 2010. Number or nuance: Which factors restrict reliable word sense annotation? In *Proceedings of LREC*.

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*.

Irina Chugur, Julio Gonzalo, and Felisa Verdejo. 2002. Polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 32–39. ACL.

Silvie Cinková, Martin Holub, and Vincent Krí. 2012. Managing uncertainty in semantic tagging. In *Proceedings of EACL*, pages 840–850. ACL.

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.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439.

J. Hong and C.F. Baker. 2011. How Good is the Crowd at "real" WSD? In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 30–37. ACL.

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

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1):15–48.

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*.

A. Kittur, B. Smus, S. Khamkar, and R.E. Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of UIST*, pages 43–52. ACM.

M. Kosinski, Y. Bachrach, G. Kasneci, J. Van-Gael, and T. Graepel. 2012. Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *ACM Web Sciences*. ACM.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Beverly Hills, CA.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, second edition.

A. Kulkarni, M. Can, and B. Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of CSCW*, pages 1003–1012. ACM.

J. J. Louviere. 1991. Best-Worst Scaling: A Model for the Largest Difference Judgments. Technical report, University of Alberta. Working Paper.

Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 17–24.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28. ACL.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of HLT*, pages 303–308. ACL.

G. Craig Murray and Rebecca Green. 2004. Lexical knowledge and human disagreement on a WSD task. *Computer Speech & Language*, 18(3):209–222.

Bryan Orme. 2009. MaxDiff Analysis: Simple Counting, Individual-Level Logit, and HB. Sawtooth Software.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *Proceedings of the Second Workshop on Scalable Natural Language Understanding Systems*. ACL.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Rebecca Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of LREC*, pages 1951–1956.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

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*.

Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012a. The MASC word sense sentence corpus. In *Proceedings of LREC*.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012b. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):209–252.

Anna Rumshisky and Olga Batiukova. 2008. Polysemy in verbs: systematic relations between senses and their effect on annotation. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 33–41. ACL.

Anna Rumshisky, Nick Botchan, Sophie Kushkuley, and James Pustejovsky. 2012. Word Sense Inventories by Non-experts. In *Procoeedings of LREC*.

Rion Snow, Brendan O'Connor, Dan Jurafsky, and Andrew Y. Ng. 2008. Cheap and fastbut is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263. ACL.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

Ng Hwee Tou, Chung Yong Lim, and Shou King Foo. 1999. A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources*.

Jean Véronis. 1998. A study of polysemy judgments and inter-annotator agreement. In *Program and advanced papers of the Senseval workshop*, pages 2–4.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2000. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of NIPS*.