New Insights from Coarse Word Sense Disambiguation in the Crowd

Adam Kapelner[1] *Krishna Kaliannan*[1] *H. Andrew Schwartz*[2]
Lyle Ungar[2] *Dean Foster*[1]
(1) The Wharton School of the University of Pennsylvania, Department of Statistics,
3730 Walnut Street, Philadelphia, PA 19104
(2) University of Pennyslvania, Department of Computer Science,
200 S. 33rd Street, Philadelphia, PA 19104
`kapelner@wharton.upenn.edu`, `kkali@wharton.upenn.edu`, `hansens@seas.upenn.edu`,
`ungar@cis.upenn.edu`, `foster@wharton.upenn.edu`

ABSTRACT

We use crowdsourcing to disambiguate 1000 words from among coarse-grained senses, the most extensive investigation to date. Ten unique participants disambiguate each example, and, using regression, we find surprising features which drive differential WSD accuracy: (a) the number of rephrasings within a sense definition is associated with higher accuracy; (b) as word frequency increases, accuracy decreases even if the number of senses is kept constant; and (c) spending more time is associated with a decrease in accuracy. We also observe that all participants are about equal in ability, practice (without feedback) does not seem to lead to improvement, and that having many participants label the same example provides a partial substitute for more expensive annotation.

KEYWORDS: Word sense disambiguation, crowdsourcing.

Proceedings of COLING 2012: Posters, pages 539–548,
COLING 2012, Mumbai, December 2012.

1 Introduction

Word sense disambiguation (WSD) is the process of identifying the meaning, or “sense,” of a word in a written context (Ide and Véronis, 1998). In his comprehensive survey, Navigli (2009) considers WSD an AI-complete problem — a task which is at least as hard as the most difficult problems in artificial intelligence. Why is WSD difficult and what is driving its difficulty? This study examines human WSD performance and tries to identify drivers of accuracy. We hope that our findings can be incorporated into future WSD systems.

To examine human WSD performance, we tap pools of anonymous untrained human labor; this is known as “crowdsourcing.” A thriving pool of crowdsourced labor is Amazon’s Mechanical Turk (MTurk), an Internet-based microtask marketplace where the workers (called “Turkers”) do simple, one-off tasks (called “human intelligence tasks” or “HITs”), for small payments. See Snow et al. (2008); Callison-Burch (2010); and Akkaya et al. (2010) for MTurk’s use in NLP, and Chandler and Kapelner (2010) and Mason and Suri (2011) for further reading on MTurk as a research platform.

We performed the first extensive look at coarse-grained WSD on MTurk. We studied a large and variegated set of words: 1,000 contextual examples of 89 distinct words annotated by 10 unique Turkers each. In the closest related literature, Snow et al. (2008) found high Turker annotation accuracy but only annotated a single word, while Passonneau et al. (2011) focused on only a few words and annotated fine-grained senses. The extensive size of our study lends itself to the discovery of new factors affecting annotator accuracy.

Our contribution is three-fold. First, we use regression to identify a variety of factors that drive accuracy such as (a) the number of rephrasings within a sense definition is associated with higher accuracy; (b) as word frequency increases, accuracy decreases even if the number of senses is kept constant; and (c) time-spent on an annotation is associated with lower accuracy. Second, we echo previous findings, mostly from non-WSD experiments, demonstrating that Turkers are respectably accurate (Snow et al., 2008), they’re approximately equal in ability (Parent, 2010; Passonneau et al., 2011), spam is virtually non-existent (Akkaya et al., 2010), responses from multiple Turkers can be pooled to achieve high quality results (Snow et al., 2008; Akkaya et al., 2010), and that workers do not improve with experience (Akkaya et al., 2010). Third, we present a system of crowdsourcing WSD boasting a throughput of about 5,000 disambiguations per day at $0.011 per annotation.

2 Methods and data collection

We selected a subset of the OntoNotes data (Hovy et al., 2006), the SemEval-2007 coarse-grained English Lexical Sample WSD task training data (Pradhan et al., 2007). The coarse-grained senses in OntoNotes address a concern that nuanced differences in sense inventories drives disagreement among annotators (Brown et al., 2010). We picked 1,000 contextual examples at random from the full set of 22,281.[1] Our sample is detailed in table 1. It consisted of 590 nouns and 410 verb examples that had between 2-15 senses each (nouns: 5.7 ± 3.0 senses, verbs: 4.7 ± 3.3 senses). For each snippet, ten annotations were completed by ten *unique* Turkers.

---

[1] We later disqualified 9 of the 1,000 because they had words with only one sense.

target word	# inst	# senses	target word	# inst	# senses	target word	# inst	# senses
affect-v	1	3	end-v	8	4	policy-n	10	3
allow-v	8	2	enjoy-v	3	2	position-n	13	7
announce-v	4	3	examine-v	3	3	power-n	12	4
approve-v	3	2	exchange-n	17	6	president-n	34	3
area-n	15	5	exist-v	1	2	produce-v	6	3
ask-v	16	6	explain-v	2	2	promise-v	3	2
attempt-v	2	2	express-v	3	3	propose-v	1	3
authority-n	3	6	feel-v	24	3	prove-v	2	6
avoid-v	2	2	find-v	7	6	raise-v	6	9
base-n	5	12	fix-v	2	6	rate-n	49	2
begin-v	7	4	future-n	16	4	recall-v	2	4
believe-v	9	2	go-v	12	14	receive-v	4	2
bill-n	18	9	hold-v	5	10	regard-v	1	3
build-v	1	4	hour-n	9	4	remember-v	8	6
buy-v	7	7	job-n	6	10	remove-v	4	2
capital-n	15	5	join-v	2	4	report-v	7	4
care-v	6	3	keep-v	11	8	rush-v	1	4
carrier-n	4	13	kill-v	5	9	say-v	104	5
chance-n	5	4	lead-v	9	7	see-v	9	10
claim-v	4	4	maintain-v	3	4	set-v	10	12
come-v	7	11	management-n	13	2	share-n	103	3
complain-v	2	2	move-n	19	4	source-n	10	6
complete-v	1	3	need-v	11	2	space-n	2	8
condition-n	7	4	network-n	9	4	start-v	8	7
defense-n	8	8	occur-v	2	4	state-n	33	4
development-n	8	3	order-n	11	9	system-n	15	7
disclose-v	5	2	part-n	14	7	turn-v	19	15
do-v	4	6	people-n	38	6	value-n	16	5
drug-n	7	3	plant-n	12	3	work-v	4	9
effect-n	7	5	point-n	27	14			

Table 1: The 89 words of the sample of 1000 OntoNotes snippets used in this study. "# inst" is the number of instances in the 1,000 with the corresponding target word. "# senses" is the number of sense choices provided by OntoNotes.

2.1 The WSD HIT

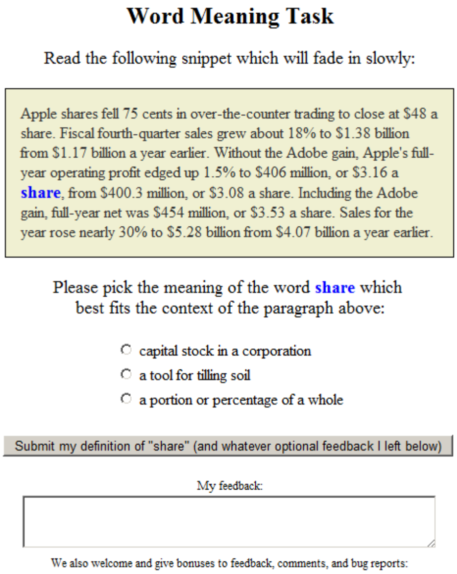
We designed a simple WSD task that was rendered inside an MTurk HIT.[2] The Turker read one example in context with the target word emboldened, and then picked the best choice from among a set of coarse-grained senses (see Figure 1). We gave a text box for soliciting optional feedback and there was a submit button below. We term a completed HIT an "annotation."

We employed anti-spam and survey bias minimizing techniques to obtain better data. We faded in each word in the context and the sense choices one-by-one at 300 words/min.[3] Additionally, we randomized the display order of the sense choices. This reduces "first response alternative bias" as explained in Krosnick (1991), but may decrease accuracy when compared to displaying

---

[2] The HIT was entitled "Tell us the best meaning of a word... do many and earn a lot! Really Easy!", the wage was $0.01, the time limit for each task was seven minutes, and the HITs expired after one hour. We posted batches of 750 new HITs to MTurk hourly upon expiration of the previous batch. Thus, the task was found readily on the homepage which drove the rapid completion.

[3] As Kapelner and Chandler (2010) found, this accomplishes three things: (1) Turkers who plan on cheating will be more likely to leave our task, (2) Turkers will spend more time on the task and, most importantly, (3) Turkers will more carefully read and concentrate on the meaning of the text.



Word Meaning Task

Read the following snippet which will fade in slowly:

Apple shares fell 75 cents in over-the-counter trading to close at $48 a share. Fiscal fourth-quarter sales grew about 18% to $1.38 billion from $1.17 billion a year earlier. Without the Adobe gain, Apple's full-year operating profit edged up 1.5% to $406 million, or $3.16 a **share**, from $400.3 million, or $3.08 a share. Including the Adobe gain, full-year net was $454 million, or $3.53 a share. Sales for the year rose nearly 30% to $5.28 billion from $4.07 billion a year earlier.

Please pick the meaning of the word **share** which best fits the context of the paragraph above:

○ capital stock in a corporation
○ a tool for tilling soil
○ a portion or percentage of a whole

Submit my definition of "share" (and whatever optional feedback I left below)
My feedback:
We also welcome and give bonuses to feedback, comments, and bug reports:

Figure 1: An example of the WSD task that appears inside an MTurk HIT. This was displayed piecewise as each word in the example ("snippet") and senses faded-in slowly.

the senses in descending frequency order as observed by Fellbaum et al. (1997). We also limited participation to US Turkers to encourage fluency in English.

Upon completion, the Turker was given an option to do another of our WSD tasks (this is the MTurk default). A Turker was not limited in the number of annotations they could do.[4] The entire study took 51 hours and cost $110. The code, raw data, and analysis scripts are available under GPL2 at `github.com/kapelner/wordturk_pilot`.

3 Results and data analysis

We were interested in investigating (1) which features in the target word, the context, and sense definition text affect Turker accuracy, (2) which characteristics in the Turker's engagement of the task affect accuracy, (3) heterogeneity in worker performance, and (4) the combination of Turker responses to boost accuracy.

We recruited 595 Turkers to work on our tasks, yielding an average accuracy of 73.4%. We measured inter-tagger agreement (ITA) using the alpha-reliability coefficient (Krippendorff, 1970) to be 0.66 (0.70 for nouns and 0.60 for verbs) which comports with Chklovski and Mihalcea (2003)'s *Open Mind Word Expert* system. However, OntoNotes was specially designed by Hovy et al. (2006) to have 90% ITA by experts. Our measure is significantly less. Untrained Turkers should not be expected to be experts.

---

[4] The actual upper limit was all 1,000 examples but in practice, not one Turker came close to completing all of them. The most productive Turker completed 405 annotations while the median completed was 4.

3.1 Performance and language characteristics

What makes WSD difficult for untrained Turkers? Are there too many senses to choose from? Is the example difficult to read? With 10,000 instances from 600 workers, we can attempt to answer these questions.

We first construct the features of interest:

- target word part-of-speech (***target word is noun?***)
- target word length in characters (***# chars in target word***)
- target word frequency (***log target word frequency***)
  log of frequency in the contemporary corpus of American English (Davies, 2008).
- number of senses to choose from (***# senses to disambiguate***)
- number of characters in the correct sense definition (***# chars in definition***)
- number of rephrasings in definition text (***# rephrasings in definition***)
  For example, the word "allot" has a sense with definition text "let, make possible, give permission" which would be counted as three rephrasings.
- number of characters in context (***# chars in context***)

We add a fixed intercept for each Turker to account for correlation among tasks completed by the same worker. The result of an ordinary least squares (OLS) regression of correct (as binary) on the variables above is presented in table 2.[5]

	estimate	$\lvert t \rvert$
target word is noun?	8.4%	7.5 ***
# chars in target word	-1.0%	3.6 ***
log target word frequency	-3.7%	7.6 ***
# senses to disambiguate	-2.9%	19.8 ***
# chars in definition	-0.063%	2.6**
# rephrasings in definition	3.4%	5.4 ***
# chars in context	-0.0062%	2.6 **

Table 2: OLS regression of instance correctness on features of the target word, context, and senses. Fixed effects for each of the 595 Turkers are not shown. ** indicates significance at the $< .01$ level, *** indicates significance at the <0.001 level.

We found that, controlling for all other variables, nouns have 8% higher disambiguation accuracy. This difference between noun and verb accuracy is also reflected in automatic system performance on the SemEval-2007 task (Pradhan et al., 2007), and often attributed to the idea that nouns "commonly denote concrete, imagible referents" (Fellbaum et al., 1997). For each extra sense, accuracy suffers 3% which also is expected since the Turkers have more choices. We show accuracy by number of senses and part of speech in figure 2. We also found the longer the target word, the more difficult the task, reflecting the fact that longer words are often more complex. Similarly, the longer the context or length of definitions decreased accuracy but the effect was quite small.

Surprisingly, with each extra rephrasing of the definition of the correct sense there is a gain of 3.5%. This suggests untrained annotators benefit from receiving a variety of sense descriptions, or that more rephrasings suggests a more coarse-grained sense which is easier for annotators to understand.

---

[5] We also ran a variety of fixed and random effects linear and logit models, all of which gave the same significance results. We chose to present the OLS output because of its familiarity and interpretability.

Finally, as the word becomes more common in the English language (controlling for all other variables, including length of word and number of senses) accuracy still suffers. Possibly the more prevalent the word in our language, the more likely it will have senses that overlap conceptually.

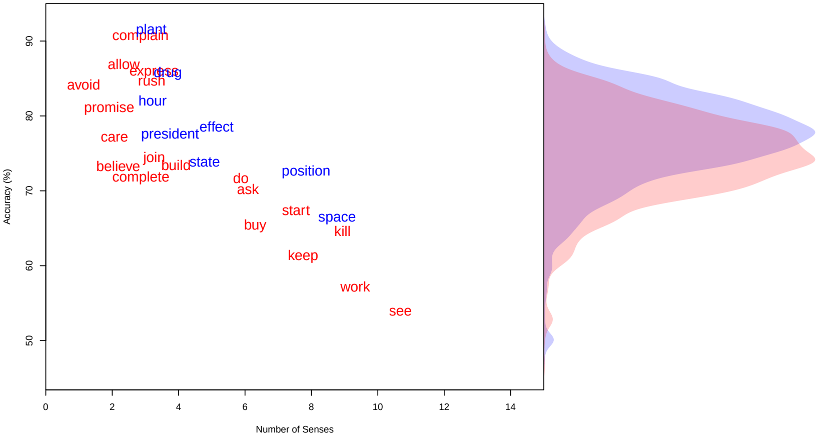


Figure 2: Predicted accuracy vs. number of senses for a sample of the words in our study. Nouns are blue; verbs are red. The densities are smoothed histograms of the noun and verb predicted accuracies. Note that the word display is jittered; there are at least two senses for each word.

3.2 Performance and Turker characteristics

Are there any characteristics about the Turker's engagement with our task that impacts accuracy? We create the following features: time spent on task, the number of words in their optional feedback message, and the number of annotations that worker completed prior to the response being examined. To control for the difficulty of each task, we added 1,000 fixed intercepts — one for each unique task; and to control for correlation among the workers, we added a fixed intercept for each worker. An ordinary least squares regression of the WSD task being correct (as binary) on the variables above[5] was run. We found that for each additional second spent on the task, accuracy drops by 0.06% ($p < 0.001$). We found that, contrary to Kapelner and Chandler (2010), leaving comments does not correspond to higher accuracy, and, in agreement with Akkaya et al. (2010), the number of tasks completed prior does not impact accuracy. This may imply that a learning effect does not exist; practice (without feedback) does *not* make perfect.

Surprisingly, spending more time on the disambiguation task associates with a significant *reduction* in accuracy ($p < 0.001$).[6] Note that this is *after* we non-parametrically control for instance difficulty and worker ability. For every additional minute spent, a Turker is 3.6% less likely to answer correctly. We posit three theories: (1) taking breaks leads to loss of

---

[6] We validated this linear approximation by regressing time spent as a polynomial and found the effect to be monotonically decreasing with a flat stretch in the middle.

concentration (2) the "knee-jerk" response is best (rumination should be discouraged), and (3) although we control for instance difficulty, an instance may only be difficult for particular workers as evidenced by their taking longer.

3.3 Turker equality

In order to replicate previous work, we investigate Turker equality and the presence of spammers and superstars via plotting the number of annotations correct by the number of annotations completed in figure 3. To test the null hypothesis that all workers are equal (and thus, average), each worker's *total contributions* are assumed to be drawn from independent Binomial random variables with probability of success $p = 73.4\%$ (the experimental average). Does the worker's confidence interval (CI) contain p? Figure 3 reveals that every worker has approximately the same capacity for performing coarse-grained WSD except for two above-average superstars and two below-average.

To test for spammers, we test against the null hypothesis of random answering, $p = 25.5\%$ (determined by simulation). Among workers who did a significant number of tasks,[7] we find only one worker who may be a spammer. We echo Akkaya et al. (2010), Snow et al. (2008), and Singh et al. (2002) and conclude there is minimal spammer contribution. Once again, we do not observe a change in accuracy by quantity of tasks completed, an observation confirmed using regression (table 3).

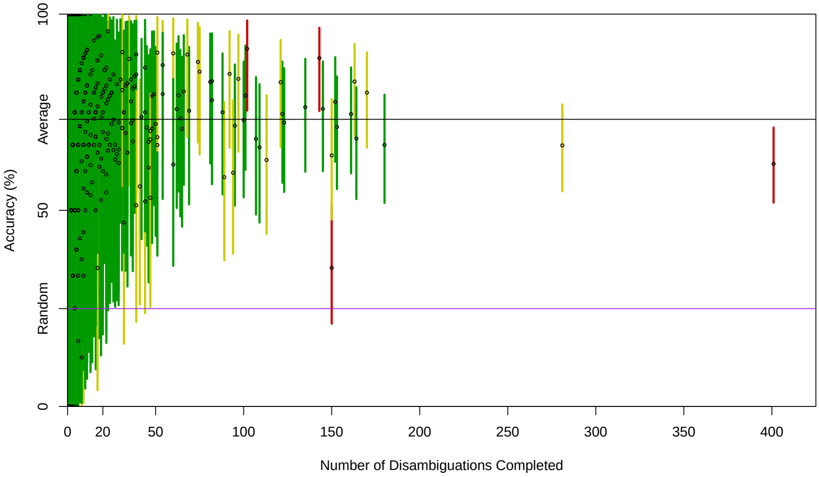


Figure 3: Accuracy of all 595 Turkers. The black line is the average accuracy ($p = 73.4\%$) and the purple line represents random sense choice accuracy (25.5%). We plot the Bonferroni-corrected Binomial proportion confidence intervals in green if they include p, yellow if the non-Bonferroni-corrected confidence intervals do not include p, and red if neither include p.

---

[7] We do not have significant power to claim a worker has accuracy of even 50% until about $n = 79$ at the Bonferroni-corrected α level.

3.4 Combining responses to optimize prediction

We can combine the 10 unique disambiguation responses for each of the 1000 examples to yield higher accuracy. Our algorithm is naive — we take the plurality vote and arbitrate ties randomly. Snow et al. (2008) found such an approach results in higher accuracy for disambiguating 'president'. We wondered if the same is true for our more extensive dataset and annotations.

# of Annotations	2	3	4	5	6	7	8	9	10	2.4 (1st plurality)
Accuracy	.734	.795	.808	.824	.830	.837	.840	.843	.857	.811

Table 3: Accuracy of the WSD task using plurality voting for different numbers of Turkers. The last column is the accuracy of the variable algorithm: starting with two workers and adding an additional worker until plurality.

Table 4 illustrates our results. There is an overall accuracy of 85.7% when annotations from all workers are aggregated. This is in the ballpark of the best supervised statistical learning techniques which boast almost 90% (Pradhan et al., 2007).[8] We determined the marginal accuracy of each added Turker by simulating random subsets of two Turkers, three Turkers, etc and employed the same plurality vote.

With techniques such as discarding results from annotators who often disagree, and giving the annotators sense choices in order of sense frequency, we could likely achieve higher accuracy.

Given MTurk annotation costs, we believe this system can be extended to accurately disambiguate a million words a year at 80% accuracy for about $25,000. This demonstrates the system's potential for mass annotation, but we reiterate that the main goal of this current work was to gain insight into drivers of WSD accuracy.

Conclusion

We performed the first extensive study of crowdsourced coarse-grained word sense disambiguation in order to gain insight into the behavioral and linguistic features that affect accuracy of the untrained annotations. As expected, we found results improved when there were less sense choices or when the target word was a noun, and that untrained workers did not improve with experience. However, we also discovered surprising insights: (1) the number of rephrasings in the correct sense definition corresponded with improved annotator accuracy, (2) frequency of target word corresponded with lower accuracy, and (3) time-spent on an individual annotation corresponded with lower accuracy. It also seems that time pressure may increase accuracy. Future experiments that prove these relationships causally may be fruitful. Lastly, we looked at Turker ability and found that they are all roughly equal in ability, and although individually not as accurate as experts, many Turkers may be pooled to improve accuracy.

Acknowledgments

We thank Mark Liberman and Lynn Selhat for helpful comments and discussions. Adam Kapelner thanks the National Science Foundation for the Graduate Research Fellowship that made this work possible.

---

[8] Note that this is not a fair comparison. These supervised algorithms were given all the training data while Turkers were *not* given any previous examples. They also arbitrated based on the senses' frequencies while we randomized the order that the senses appeared in. Finally, they were not limited to polysemous words as we were.

References

Akkaya, C., Conrad, A., and Wiebe, J. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203.

Brown, S. W., Rood, T., and Palmer, M. (2010). Number or nuance: Which factors restrict reliable word sense annotation? In *LREC: The International Conference on Language Resources and Evaluation*, pages 3237–3243.

Callison-Burch, C. (2010). Creating speech and language data with Amazon's Mechanical Turk. *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Chandler, D. and Kapelner, A. (2010). Breaking monotony with meaning: Motivation in crowdsourcing markets. *University of Chicago mimeo*.

Chklovski, T. and Mihalcea, R. (2003). Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of the Conference on Recent Advances on Natural Language Processing*.

Davies, M. (2008). *The Corpus of Contemporary American English: 425 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.*

Fellbaum, C., Grabowski, J., Landes, S., and L, S. (1997). Analysis of a hand-tagging task. In *Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics*, pages 34–40.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90 percent solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60. Association for Computational Linguistics.

Ide, N. and Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):2–40.

Kapelner, A. and Chandler, D. (2010). Preventing Satisficing in Online Surveys: A "Kapcha" to Ensure Higher Quality Data. In *CrowdConf ACM Proceedings*.

Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Mason, W. and Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods, Forthcoming*.

Navigli, R. (2009). Word sense disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1–69.

Parent, G. (2010). Clustering dictionary definitions using Amazon Mechanical Turk. *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 21–29.

Passonneau, R. J., Bhardwaj, V., Salleb-Aouissi, A., and Ide, N. (2011). Multiplicity andword sense: Evaluating and learning from multiply labeledword sense annotations. *Language Resources and Evaluation*.

Pradhan, S., Loper, E., and Dligach, D. (2007). Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92.

Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., and Li Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Morristown, NJ, USA. Association for Computational Linguistics.