Crowdsourcing Preposition Sense Disambiguation with High Precision via a Priming Task

Shira Wein
Georgetown University
sw1158@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

The careful design of a crowdsourcing protocol is critical to eliciting highly accurate annotations from untrained workers. In this work, we explore the development of crowdsourcing protocols for a challenging word sense disambiguation task. We find that (a) selecting a similar example usage can serve as a proxy for selecting an explicit definition of the sense, and (b) priming workers with an additional, related task within the HIT improves performance on the main proxy task. Ultimately, we demonstrate the usefulness of our crowdsourcing elicitation technique as an effective alternative to previously investigated training strategies, which can be used if agreement on a challenging task is low.

1 Introduction

Crowdsourcing work relies on effective protocol design in order to elicit meaningful and accurate annotations from lay workers. Traditional crowdsourcing protocols for challenging tasks train crowd workers, incentivize high quality work with bonuses, and filter out workers with pre-determined gold annotations.

In this work, we demonstrate a novel crowdsourcing technique which frames the task such that the worker is able to provide a high quality annotation for challenging tasks without extensive prior training. We consider the challenging task of English preposition supersense disambiguation.

Prepositions are difficult to disambiguate due to their extreme polysemy and frequency (Litkowski and Hargraves, 2007; Hovy et al., 2010; Gong et al., 2018). While some distinctions—such as the locative vs. temporal ambiguity of ***in*** *Seattle* vs. ***in*** *October*—are quite intuitive, the semantic range of English prepositions makes fully disambiguating them a daunting task. Consider the sentence "Nice and quiet place **with**$_{PartPortion}$ cosy living room just outside the city." Though **with** here denotes a part of a whole (a room in an abode), it is also describing a characteristic of the place/living room, suggesting that **with**$_{Characteristic}$ would also be a reasonable annotation. With this being a difficult task because of the inherently categorical nature of word sense disambiguation, we aim to develop simple linguistic tasks which elicit annotations by proxy.

Prepositions are critical to language understanding (Kim et al., 2019), which makes preposition sense disambiguation an important task, impacting a range of downstream applications, including: relation extraction (Elazar et al., 2021), paraphrasing of phrasal verbs (Gong et al., 2018) and noun compounds (Ponkiya et al., 2018; Hendrickx et al., 2013), machine translation (Parameswarappa and Narayana, 2012; Chiang et al., 2009; Chan et al., 2007), semantic role labelling (Ye and Baldwin, 2006; Srikumar and Roth, 2011), and more. Automatic supersense classifiers exist for preposition sense disambiguation (Liu et al., 2021), but are unable to achieve the precision levels of trained annotators.

In order to tackle the challenging aim of preposition sense annotation, we design linguistic tasks to enable the elicitation of high-quality annotations. We demonstrate that a weak form of guidance, which we call a *priming task*, improves performance on the main task.

In our case, the main task asks the worker to select the most similar usage to the prompt. However, the notion of meaning similarity between preposition usages may not be apparent to crowd workers. We find that first asking the worker to select an appropriate definition is an effective priming task: its mere presence enhances workers' ability to choose the correct exemplar in the second question. This related (priming) task serves as a high-precision, low-overhead alternative to training.

Finally, we compare the predictions of the crowd with those of a classifier, and find that they are largely complementary; for most prepositions

tested, combining the two predictions gives perfect or near-perfect precision.

Related work has proposed the ability to crowdsource supersenses and tested a pilot on trained in-house annotators (Gessler et al., 2020); in this work, we develop new formulations and collect actual crowdsourced annotations from Mechanical Turk to ascertain the suitability of our formulations.

Our contributions include:

- a **novel crowdsourcing elicitation technique** which primes implicit questions about lexical meaning with an explicit task
- a promising crowdsourcing **approach to preposition supersense annotation** to elicit high-precision labels for a subset of instances without prior annotator training
- a demonstration that the crowd and a supersense **classifier** give complementary signals conducive to ensembling.

2 Related Work on Crowdsourcing

The Amazon Mechanical Turk online crowdsourcing platform enables crowd workers from around the world to perform short tasks published by Requesters. Mechanical Turk has been used for a range of annotation tasks.

Mechanical Turk is used in many cases where a large amount of annotations are collected. If these many annotations are of lesser quality, it is not necessarily the case that having more data is better than having a small amount of expert annotations would have been (Bhardwaj et al., 2010).

Crowdsourcing Protocols. A variety of alternatives or additions to worker training have been employed in previous work. Notably, qualifying workers (filtering out who can and cannot complete the task) is a method akin to training which seeks to ensure higher quality annotations. Recent work evaluating crowdsourcing protocols for the development of natural language understanding datasets ask crowd workers to write a multiple choice question with one of four qualification follow ups; with the goal of making more difficult NLU questions, this work finds that training crowd workers, sending feedback, and qualifying crowd workers is an effective strategy (Nangia et al., 2021). Qualifying workers beforehand based on self-assessment is flawed due to bias in the self-assessment; leveraging a combination of self-assessment and performance on the task is a more useful filtering process (Gadiraju et al., 2017).

Lee et al. (2022) provides an ordering strategy for sentence-level annotation tasks such that the annotators learn and improve from task to task. Mikulová et al. (2022) considers the effect of pre-annotation and task design on dependency syntax annotation, finding that automatic pre-annotation is useful while other support tools are not as beneficial.

The *retainer method* has been developed for real-time crowdsourcing, using retainer pools (qualifying workers into a group of candidates who can be called upon quickly to complete the task), push notifications, and recruiting workers before the task is actually published (Bernstein et al., 2012). Vaughan (2017) put forward a set of best practices for Requesters, which includes providing clear instructions and iteratively piloting task designs.

The idea of "priming" a crowd worker to influence their performance (through the design of instructions, the order of items, etc.) has been explored in previous work. Jiang et al. (2017) investigated how the choice of provided examples and the phrasing of the prompt affect worker response. Federmann et al. (2019) further investigated the effect of prompt question on worker performance. Additional work also explored how providing surrounding text (Mitchell et al., 2014), drawings as prompts (Kumaran et al., 2014), lists of options (Wang et al., 2012), and diverse word suggestions (Yaghoub-Zadeh-Fard et al., 2020) improve worker performance. Jiang et al. (2018) suggests that workers are cognitively "primed" to replicate their own mistakes when completing numerous HITs in sequence. Colombini (2018) similarly investigates "inter-task effect" (effect from one HIT to the next), finding that consistency in task structure across HITs improves worker performance.

We introduce the concept of a *priming task*, an additional task designed to be completed first in order to improve performance on a related *main task*. The idea is that the priming task is a less ambiguous task not intended to collect data, but rather to frame the worker's thinking so as to draw attention to the relevant aspects of the main task.[1]

Word sense disambiguation is a difficult task (Artstein and Poesio, 2008), which is particularly true for prepositions, as they are widely polysemous (Gong et al., 2018; Hovy et al., 2010). Prior work has proposed, but not executed, the possibility

[1] The 2 tasks are questions that appear together on the same screen (in the same HIT), and the main task is actually a proxy task because it is an indirect form of labeling.

of crowdsourcing preposition sense disambiguation for the supersense schema—Gessler et al. (2020) proposed that, instead of training annotators to apply abstract labels, like STARTTIME, GESTALT, and AGENT, annotators could be asked to perform simpler *proxy tasks* from which the supersense labels could be inferred automatically. While Gessler et al. conducted a pilot with trained in-house annotators, we introduce a new definition-based formulation, which we combine with an exemplar-based approach, and collect annotations from Mechanical Turk to establish whether sense disambiguation of prepositions is able to be crowdsourced in practice.

3 Crowdsourcing Protocol

3.1 Task

Our crowdsourcing task is the challenging task of preposition sense disambiguation (§1), via the (indirect) annotation of *supersenses*. [2] We want to determine whether the most frequent senses of prepositions can be categorized by the crowd, such that the majority of prepositions could be annotated at a large scale, and any remaining long-tail cases could then by annotated by experts.

In this work, we ask: can we elicit preposition supersense data from crowd workers with high precision? Because the supersense training process is extensive and specialized, we leverage a *proxy task*, through which the crowd judgments can be converted into actual supersense labels. As a result of our study, we propose the technique of a *priming task*, which is designed to improve performance on the main task rather than to collect data.

The SNACS *supersense* annotation schema and corpus extensively document the senses of English prepositions and possessives, with 50 supersense classes categorizing the use of an adposition in context (Schneider et al., 2018). While some distinctions—such as the locative vs. temporal ambiguity of ***in*** *Seattle* (supersense label LOCUS) vs. ***in*** *October* (TIME)—are quite intuitive, the semantic versatility of prepositions requires extensive annotator training. Preposition supersenses form a subset of labels in the lexical semantic recognition task Liu et al. (2021).

From the STREUSLE corpus (Schneider and Smith, 2015), we use annotated tokens of 6 preposition types: **from**, **in**, **on**, **with**, **for**, and **of**, which we choose due to their high polysemy and frequency, collectively comprising more than 60% of all preposition tokens in the STREUSLE corpus. The gold annotations are used to evaluate our approach.

[2] Technically two supersense fields are annotated per token; they may be the same or different (Schneider et al., 2018).

3.2 Two Task Designs

We evaluate crowd predictions with the aim of prioritizing precision over recall, because we want to ensure that any sense that receives an annotation by majority vote is very likely to be accurate. Our first task design provides definitions and examples of possible senses for the preposition, drawing from the traditional notion of word sense disambiguation being obtained by selecting a sense from a list of definitions as the *priming* task (cf. Ahlswede and Lorand, 1993; Jurgens, 2013; Tratz, 2011). The second design uses BERT (Devlin et al., 2019) to retrieve nearest neighbors from a gold-annotated seed corpus as the *proxy* task.

Definition-Based The Definition-Based approach presents the *target sentence* (the sentence to be annotated), and asks a question about the relationship between the governor and object of the preposition. The question presented to the crowd worker is "The word [preposition] expresses a relationship between two things. Which of the following options, if any, describes the kind of relationship?" The *definitions* include a simple, short description of what the preposition may be conveying in that sentence, as well as examples of that usage. We wrote a small number of definitions for each of the 6 prepositions, mindful to avoid jargon and to avoid imposing a high cognitive load with many options. These are presented as options along with "None of the above" and "Not sure/sentence is hard to understand". Since only a few of the possible senses receive definitions, annotators are encouraged to select "None of the above" if none is a good match. For example, for the relationship options included in figure 1, the four options provided cover 75% of the STREUSLE instances of the preposition **from**, corresponding respectively to the annotations: ORIGINATOR⇝SOURCE, SOURCE⇝SOURCE, LOCUS⇝SOURCE, and STARTTIME⇝STARTTIME, the last of which is the correct choice.

Exemplar Matching Our Exemplar Matching task utilizes contextualized embeddings from BERT (Devlin et al., 2019) to identify n nearest neighbors of the use of a preposition in a sentence. The intuition behind the exemplar-based approach is that annotators would have an easier time identi-

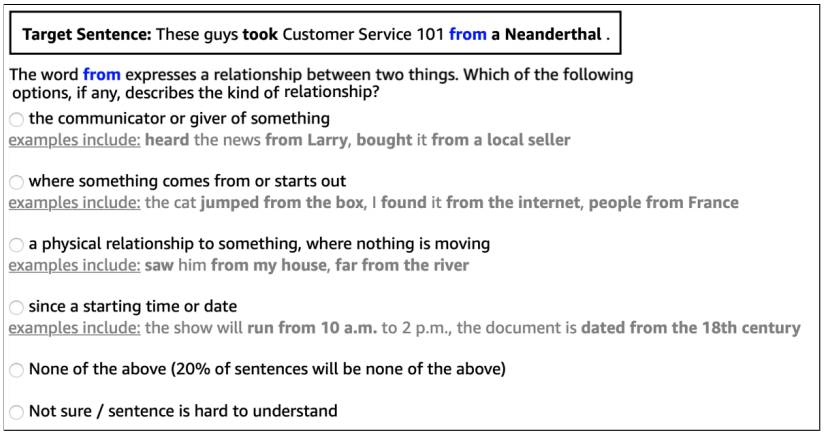

Figure 1: Definition-Based approach for a sentence from the test set in the STREUSLE dataset.

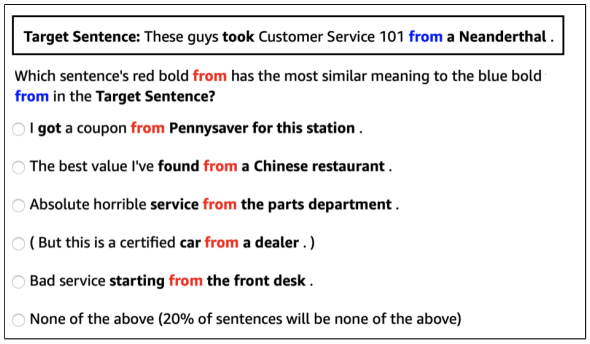

Figure 2: Exemplar Matching approach for the same STREUSLE test set sentence as in figure 1, using the nearest neighbors from the STREUSLE train set as identified by BERT.

fying a similar use of a preposition in context than explicitly describing the use of the preposition.

STREUSLE is split into train, test, and dev sets. We use pre-trained (without any fine-tuning) BERT-base-uncased to collect a given test or dev sentence's 5 nearest unique neighbors in the train set, using cosine similarity. We concatenate the vectors from the last four layers, and only use the indexed embedding of the preposition in question to find nearest neighbors. We retrieve the 5 nearest neighbors with distinct supersenses, and present these 5 options plus a "None of the above" option.

Subject to the performance of the nearest neighbor retrieval metric, oracle recall of the crowd workers is therefore 87% overall, i.e. for our experiments, 87% of the retrieved nearest neighbors include a gold label in the 5 unique options. We also found that ensuring that the 5 options provided reflected unique supersenses did increase the likelihood that the majority consensus would match that of the gold label.

An example of this protocol can be seen in figure 2, which features the same sentence as figure 1. Figure 2 demonstrates the **Exemplar Matching** proxy task, with the 5 nearest neighbor preposition usages as well as a "None of the above" option being presented to the worker.

Combined In this design, we include in each HIT the Definition-Based task followed by the Exemplar Matching task, for the same sentence. Within the same HIT, the Definition-Based task is first presented and then below that (seen by the crowd worker by scrolling down) the Exemplar Matching task is shown.

3.3 Mechanical Turk Experimental Setup

Using the Mechanical Turk crowdsourcing platform, we have 5 crowd workers annotate each sentence. With there being 160 sentences (130 unique sentences, as the **for** experiment was performed twice with disjoint sets of annotators), we collect 1,650 judgments total. Consensus is established when 3 of 5 crowd workers select the same option—a simple majority balances precision and recall.

We filter the target and exemplar sentences to select instances of canonical phrase order between the syntactic governor, the preposition, and the object, such that the governor precedes (not necessarily immediately) the preposition, which also precedes the object (again, not necessarily immediately).[3] Target sentences were sampled randomly from this filtered set.

[3]This filtered set includes approximately 44% of the instances of the selected preposition types. We filter the data for simplicity of interpretation of results, and the filtered task would not be a more or less difficult task for annotation.

			Exemplar Matching			*Combined: EM*			*Combined: DB*			
Prep.	*n*	**Classifier P**	**P**	**R**	**O R**	**P**	**R**	**O R**	**P**	**R**	**O R**	**Ensemble P**
for	30	0.70	0.76	0.53	0.90	**0.90**	0.63	0.90	0.74	0.57	0.57	18/19=0.95
for *(rpt.)*	30	0.70	0.71	0.50	0.90	**0.77**	0.57	0.90	0.74	0.57	0.57	15/15=1.00
of	20	**0.75**	0.33	0.20	0.90	0.42	0.25	0.90	0.29	0.10	0.40	4/4=1.00
from	10	0.50	0.60	0.30	0.80	0.67	0.40	0.80	**0.71**	0.50	0.80	2/2=1.00
in	30	0.73	**0.92**	0.77	0.87	0.88	0.73	0.87	0.88	0.73	0.80	22/22=1.00
on	15	0.60	0.83	0.33	0.93	**1.00**	0.60	0.93	0.92	0.80	0.87	6/6=1.00
with	30	0.43	0.30	0.10	0.80	0.47	0.27	0.80	**0.53**	0.30	0.43	7/10=0.70

Table 1: The results of the Exemplar Matching (EM) task alone, the Definition-Based (DB) task within the Combined approach, and the EM task within the Combined approach. For each preposition, the same *n* sentences are used in the EM and the Combined HITs, where *n* indicates the number of sentences tested for that preposition. Recall (R) is out of all *n*, not only the sentences which have gold options. This is also reflected in the Oracle Recall (O R) which indicates the best possible crowd performance given that not all tasks present a gold option. Precision is labeled as P. The rightmost column shows Ensemble results, discussed in §4.2.

4 Results & Analysis

Before coming to the task design presented in this work, we iterated over many approaches and techniques to crowdsourcing preposition sense disambiguation. In total, we elicited 4,080 annotations and developed 17 slight variations of the approach presented here, which primes the crowd workers by combining the definition-based and exemplar-based tasks. We saw consistent trends across these pre-study variations, resulting in the current approaches (presented in this work), which prime the crowd workers by combining the definition-based and exemplar-based tasks. For these results, we collect between 10 and 30 annotations for each of the 6 prepositions, via two methods: the Exemplar Matching design alone, and the Combined design.

4.1 Effect of Task Design

The precision and recall scores from the different task designs appear in table 1, accompanied by a classifier baseline. Precision is important for this task, because we want to have confidence that the annotations that are being produced are trustworthy. For 5 of the 6 prepositions tried, the best crowd design outperforms the classifier; notably, the precision for 3 prepositions—**for**, **in**, and **on**—is consistently high. The biggest exception is **of**, which is extremely polysemous, and the classifier picks up on the plethora of options with higher precision than the crowd workers (at least in our small sample). **With** achieves better precision from the Combined judgments than the classifier, though it barely scratches 50%.

We see that within the Combined design, the Exemplar Matching judgments are equal or superior to the Definition-Based judgments in all experiments except for **from**. The two-step process takes slightly longer for annotators and thus is slightly more expensive, but achieves very high precision, which is the aim here. Notably, the Exemplar Matching judgments are usually enhanced by the presence of the Definition-Based task in the HIT: Combined EM precision surpasses plain Exemplar Matching in all but the **in** experiment. This is surprising because the combined approach is useful even when the Definition-Based options do not include a correct definition, or when the annotator doesn't choose the correct definition—the presence of the task alone results in a statistically significant improvement in precision.The improved precision for Combined EM vs. plain EM is statistically significant per a *t*-test: Paired Two Sample for Means on precision resulting in a one-tailed *p*-value of 0.0083.

This suggests that the definitions are biasing crowd workers to attend to the preposition's meaning in context, *even if they don't actually choose the correct definition*. This insight may be useful for other kinds of crowdsourcing tasks, such as other word sense disambiguation tasks, which currently rely on only one annotation elicitation method, but could benefit from a combination of various methods. In particular, **if inter-annotator agreement is low on a more challenging task, a specialized protocol could be used to prime the crowd workers' thinking, without training.** While it is well-known that initial questions in a survey can bias answers to subsequent questions (Schuman, 2008), we are not aware of other crowdsourcing studies that have exploited this to improve workers' performance on a task by including another task first. We have demonstrated that this approach is successful on the challenging task of preposition sense disambiguation; future work should explore

the utility of similar schemes for other tasks.

We ran the **for** experiment twice with the same instances to see how much randomness in the annotator sample would affect results. The absolute scores in some conditions were slightly affected, but in both cases the Combined: EM design was the best and all crowd designs outperformed the classifier.

4.2 Ensemble Results

In comparison to the highest performing supersense tagger (Liu et al., 2021), our crowdsourcing approach generally achieves higher precision. However, we find that ensembling (predicting only when the crowd workers via the Combined EM approach agrees with the supersense classifier) can give extremely high-precision predictions, showing that the two signals are complementary.

Specifically, we consider an ensemble where an annotation is produced only if the majority of crowd workers in the best design (Combined Exemplar Matching) and the classifier are in agreement. For the ensemble, the precision is perfect or near-perfect for 5 of 6 prepositions, as seen in the rightmost column of table 1 (with the caveat that counts for some prepositions are too low to draw firm conclusions). The senses annotated via crowdsourcing alone are more varied than the senses annotated by the ensemble approach, but the ensemble approach is more reliable for more frequent senses. If we extrapolate the highly effective ensemble technique to the entirety of the relatively small STREUSLE corpus, the Combined Exemplar Matching approach would result in an estimated 808 annotations (approx. 606 of which would be correct), while the ensemble would produce an estimated 563 annotations (approx. 534 correct). This gives a sense of the potential for efficiency gains when applied on a larger scale.

Note that though the precision will be extremely high, to ensemble the crowdsourcing approach and classifier means that slightly fewer annotations will be produced. Of the 2,692 STREUSLE instances of the 6 prepositions used in this study, 1,191 of them meet the same filtering requirement we used (such that the governor precedes the preposition, which precedes the object). Extrapolating to these 1,191 tokens, the Combined Exemplar Matching approach would result in an estimated 808 annotations (an estimated 606 of which would be correct), while the ensemble would produce an estimated 563 annotations (an estimated 534 of which would be correct).

5 Conclusion

This paper outlined a promising approach to crowdsourcing preposition supersense annotation (a particularly challenging form of word sense disambiguation). The crowd workers outperformed automatic supersense tagging on 5 of the 6 prepositions studied. We compared multiple designs, finding that prompting annotators to reason both explicitly and implicitly about meaning is most effective, even when the explicit question does not elicit a correct annotation. The crowdsourcing approach achieves very high precision and acceptable recall for 3 prepositions.

6 Ethics

When using a crowdsourcing platform like Mechanical Turk, it is critical to ensure fair payment and treatment of crowd workers. For the Exemplar Matching tasks alone, we paid $0.15 per HIT, and for the combined Definition-Based and Exemplar Matching task, we paid $0.20 per HIT. Per reviews on the TurkerView website, which Turkers use to anonymously review Requesters, the reviews are positive, indicating that the payment is approved quickly, with Turkers never being rejected or blocked. The average hourly wage is reported on the site based on the payment for completion of the task and how long it takes to complete the task (as reported by the worker, rather than the time spent between accepting the task and submitting, as reported by Mechanical Turk), and reflects an average hourly wage of $20.66 based on 40 reports. IRB exemption was granted for this study. Intending to have this task primarily completed by native English speakers, we filtered crowd workers to require that the Location is United States and the number of approved HITS is greater than 5000.

Acknowledgements

Thank you to Vivek Srikumar and anonymous reviewers for their feedback. This work is supported by a Clare Boothe Luce Scholarship.

References

Thomas Ahlswede and David Lorand. 1993. Word sense disambiguation by human subjects: Computational and psycholinguistic applications. In *Acquisition of Lexical Knowledge from Text*.

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

Michael S. Bernstein, David R. Karger, Robert C. Miller, and Joel Brandt. 2012. Analytic methods for optimizing realtime crowdsourcing. *arXiv:1204.2995 [physics]*. ArXiv: 1204.2995.

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55, Uppsala, Sweden. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic. Association for Computational Linguistics.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, Colorado. Association for Computational Linguistics.

Brian J Colombini. 2018. Worker retention, response quality, and diversity in microtask crowdsourcing: An experimental investigation of the potential for priming effects to promote project goals.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2021. Text-based np enrichment.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China. Association for Computational Linguistics.

Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based preselection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction*, 24:1–26.

Luke Gessler, Shira Wein, and Nathan Schneider. 2020. Supersense and sensibility: Proxy tasks for semantic annotation of prepositions. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 117–126, Barcelona, Spain. Association for Computational Linguistics.

Hongyu Gong, Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. Preposition sense disambiguation and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1510–1521, Brussels, Belgium. Association for Computational Linguistics.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a preposition? dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Posters*, pages 454–462, Beijing, China. Coling 2010 Organizing Committee.

Youxuan Jiang, Catherine Finegan-Dollak, Jonathan K. Kummerfeld, and Walter Lasecki. 2018. Effective crowdsourcing for a new type of summarization task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 628–633, New Orleans, Louisiana. Association for Computational Linguistics.

Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding task design trade-offs in crowdsourced paraphrase collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 103–109, Vancouver, Canada. Association for Computational Linguistics.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, Georgia. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Kumaran, Melissa Densmore, and Shaishav Kumar. 2014. Online gaming for crowd-sourcing phrase-equivalents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1238–1247, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. 2022. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373.

Ken Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24–29, Prague, Czech Republic. Association for Computational Linguistics.

Nelson F. Liu, Daniel Hershcovich, Michael Kranzlein, and Nathan Schneider. 2021. Lexical semantic recognition. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 49–56, Online. Association for Computational Linguistics.

Marie Mikulová, Milan Straka, Jan Štepánek, Barbora Štepánková, and Jan Hajic. 2022. Quality and efficiency of manual annotation: Pre-annotation bias.

Margaret Mitchell, Dan Bohus, and Ece Kamar. 2014. Crowdsourcing language generation templates for dialogue systems. In *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, pages 172–180, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult nlu data collection tasks?

S. Parameswarappa and V.N. Narayana. 2012. Sense disambiguation of simple prepositions in english to kannada machine translation. In *2012 International Conference on Data Science Engineering (ICDSE)*, pages 203–208.

Girishkumar Ponkiya, Kevin Patel, Pushpak Bhattacharyya, and Girish Palshikar. 2018. Treat us like the sequences we are: Prepositional paraphrasing of noun compounds using LSTM. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1827–1836, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Howard Schuman. 2008. *Method and meaning in polls and surveys*. Harvard University Press.

Vivek Srikumar and Dan Roth. 2011. A joint model for extended semantic role labeling. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 129–139, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. Ph.D. dissertation, University of Southern California, Los Angeles, California.

Jennifer Wortman Vaughan. 2017. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18:193:1–193:46.

William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 73–78.

Mohammad-Ali Yaghoub-Zadeh-Fard, Boualem Benatallah, Fabio Casati, Moshe Chai Barukh, and Shayan Zamanirad. 2020. Dynamic word recommendation to obtain diverse crowdsourced paraphrases of user utterances. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 55–66.

Patrick Ye and Timothy Baldwin. 2006. Semantic role labeling of prepositional phrases. *ACM Transactions on Asian Language Information Processing*, 5(3):228–244.