Scaling Semantic Frame Annotation

Nancy Chang, Google, ncchang@google.com
Praveen Paritosh, Google, pkp@google.com
David Huynh, Google, dfhuynh@google.com
Collin F. Baker, ICSI, collinb@icsi.berkeley.edu

Abstract

Large-scale data resources needed for progress toward natural language understanding are not yet widely available and typically require considerable expense and expertise to create. This paper addresses the problem of developing scalable approaches to annotating **semantic frames** and explores the viability of crowdsourcing for the task of **frame disambiguation**. We present a novel **supervised crowdsourcing** paradigm that incorporates insights from human computation research designed to accommodate the relative complexity of the task, such as exemplars and real-time feedback. We show that non-experts can be trained to perform accurate frame disambiguation, and can even identify errors in gold data used as the training exemplars. Results demonstrate the efficacy of this paradigm for semantic annotation requiring an intermediate level of expertise.

1 The semantic bottleneck

Behind every great success in speech and language lies a great corpus—or at least a very large one. Advances in speech recognition, machine translation and syntactic parsing can be traced to the availability of large-scale annotated resources (Wall Street Journal, Europarl and Penn Treebank, respectively) providing crucial supervised input to statistically learned models.

Semantically annotated resources have been comparatively harder to come by: representing meaning poses myriad philosophical, theoretical and practical challenges, particularly for general purpose resources that can be applied to diverse domains. If these challenges can be addressed, however, semantic resources hold significant potential for fueling progress beyond shallow syntax and toward deeper language understanding.

This paper explores the feasibility of developing scalable methodologies for semantic annotation, inspired by three strands of work.

First, **frame semantics**, and its instantiation in the Berkeley **FrameNet** project (Fillmore and Baker, 2010), offers a principled approach to representing meaning. FrameNet is a lexicographic resource that captures syntactic and semantic generalizations that go beyond surface form and part of speech, famously including the relationships among words like *buy*, *sell*, *purchase* and *price*. These rich structural relations provide an attractive foundation for work in deeper natural language understanding and inference, as attested by the breadth of applications at the Workshop in Honor of Chuck Fillmore at ACL 2014 (Petruck and de Melo, 2014). But FrameNet was not designed to support scalable language technologies; indeed, it is perhaps a paradigm example of a hand-curated knowledge resource, one that has required significant expertise, training, time and expense to create and that remains under development.

Second, the task of **automatic semantic role labeling (ASRL)** (Gildea and Jurafsky, 2002) serves as an applied counterpart to the ideas of frame semantics. Recent progress has demonstrated the viability of training automated models using frame-annotated data (Das et al., 2013; Das et al., 2010; Johansson and Nugues, 2006). Results based on FrameNet data have been limited by its incomplete

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 1–10,
Denver, Colorado, June 5, 2015.

lexical coverage (since the project is ongoing) as well as the relatively limited amount of annotated data. More impressive results have been based on PropBank (Palmer et al., 2005), a semantic resource whose frames are more lexically specific than those of FrameNet. PropBank frames are generally more tightly linked to surface syntax (and thus afford less generalization across words), but are relatively simpler to define and annotate, as reflected by its greater coverage and amount of annotated data. It seems natural to investigate whether a comparable amount of FrameNet data would yield equally good performance (along with the further benefits of frame-level generalizations).

Third, a handful of studies from the relatively new field of **human computation** suggest that some aspects of frame annotation may be amenable to non-expert curation, such as made possible by crowdsourcing platforms like Amazon Mechanical Turk (AMT) (Hong and Baker, 2011; Fossati et al., 2013). These findings are not altogether surprising: frame semantics purports to capture generalizations that depend on everyday, non-specialist language use. Frame annotation should therefore not require the same level of training as, for example, syntactic annotation. On the other hand, while competent speakers of a language are assumed to make implicit use of frame-like structures—i.e., understanding who did what to whom and other kinds of relationships implied by a specific expression—they do not explicitly annotate semantic information as a natural part of everyday language use. Thus, unlike translation—which (some) humans do rather naturally—frame annotation is unlikely to occur in the wild, and will likely require more instruction than a typical AMT task.

These three strands together suggest that frame semantics is a promising option for meaning representation; that larger-scale frame-annotated data could drive ASRL models; and that the task of frame annotation may be amenable to crowdsourcing methods. We take these strands as a starting point for exploring how richer human computation frameworks can support scalable frame annotation, focusing in this paper on one part of frame annotation (the **frame disambiguation** task).

In the remainder of the paper, we first describe relevant previous work in more detail (Section 2). We then introduce a novel **supervised crowdsourcing** framework that adapts previous work by introducing multiple kinds of feedback and supervision (Section 3) and describe experiments using this framework to crowdsource frame disambiguation (Section 4). Finally, we discuss results and future avenues suggested by this research (Section 5), in particular the possibility that non-experts can be efficiently and effectively trained to perform tasks requiring an intermediate level of expertise.

2 Background

In this section we briefly describe the target representation of semantic frames, the FrameNet resource, the frame disambiguation annotation task, and some relevant past human computation efforts.

2.1 Frame semantics

A **semantic frame** (or simply **frame**), as developed by the late Charles J. Fillmore (Fillmore, 1976; Fillmore, 1982), is a conceptual gestalt that represents a generalization over similar scenes—typically corresponding to events, relations, states, or entities. Frames are structured around a set of semantic **roles**, also called **frame elements** (FEs), corresponding to participants in the scene.

The key theoretical insight of **frame semantics** is that the meanings of most words (and other constructions) can be understood in relation to the semantic frames they evoke. The much-discussed Commercial Transaction frame, for example, has FEs for the Buyer, Seller, Goods and Money; and it is associated with a set of words, or **lexical units** (LUs), that **profile** (or highlight) different FEs or sets of FEs (e.g., the verb *buy* is typically expressed along with the Buyer and the Goods FEs, while the noun *price* is mainly associated with the Money).

Frames vary considerably in complexity and level of granularity. Moreover, individual **lemmas** (or words) might be associated with multiple frames. For example, the lemma *like* (as a preposition and verb, respectively) is associated with two frames:

- Similarity: *Skiing is LIKE windsurfing.*
- Experiencer focus: *I LIKE looking in windows.*

The same lemma with the same part of speech can also be ambiguous, as in the case of *century*:

- Measure duration: *CENTURIES of farming have shaped our countryside.*
- Calendric unit: *By the 13th CENTURY....*

For simplicity, the examples above do not show the FEs defined for each frame and how they relate to different parts of the text, but a fully frame-annotated sentence would include that information.

2.2 FrameNet and frame disambiguation

FrameNet is a lexical resource for English based on frame semantics, in development since 1997 (Fillmore and Baker, 2010; Ruppenhofer et al., 2006). It includes nearly 1,200 frame definitions; 200,000 manually annotated examples; and about 13,000 LUs linked to specific frames.

The frame annotation process traditionally employed by Berkeley FrameNet combines **frame creation** with **lexicographic frame annotation**, where annotators select sentences from a corpus containing a lemma illustrating a frame. A separate **full-text frame annotation** process attempts to annotate all frames evoked by a sentence.

For either style of frame annotation, one must decide whether a lemma used in a given sentence is an instance of a particular frame, or more generally decide which of several candidate frames it evokes. Since the FrameNet project is ongoing (i.e., many frames have not yet been defined), the evoked frame may not even be among the known candidate frames. We call this task **frame disambiguation** (FD), corresponding roughly to word sense disambiguation.

FD is only the first step toward complete frame semantic annotation. The second is **frame element annotation** (FEA), the assignment of FEs to words in the sentence. The output of FEA corresponds to that of ASRL systems like those mentioned above; these systems often make precisely the same division of labor among FD and FEA phases (Das et al., 2013).

2.3 Insights from human computation

Human computation, in particular the use of large numbers of non-expert judgments to complement or substitute for expert judgments, has been well-established for many types of data collection, both commercial and scientific. Several crowdsourcing experiments have explored frame disambiguation and related tasks.

2.3.1 Crowdsourcing for frame disambiguation

The most relevant precursor of the current work is a series of experiments on crowdsourcing frame annotation, in particular the frame disambiguation task, using Amazon Mechanical Turk (AMT), reported at LAW V (Hong and Baker, 2011).

The target sentences consisted of unannotated sentences from the FrameNet database, plus a few annotated sentences for measuring annotator accuracy. Several task designs were tried:

- frame choice: Workers choose from a list of candidate frames, plus "None of the above".
- simplified frame names: as above, but with FrameNet terms rewritten for non-experts.
- frame sorting, with randomly chosen gold exemplars: Workers see a list of sentences and "piles" corresponding to candidate frames, each with a starter gold exemplar. They sort sentences into the appropriate frame pile (and freely recategorize sentences if desired).

Several experiments were run with the last design, varying the qualifications of the workers and the pay rate, over words with varying degrees of ambiguity.

The results showed that AMT workers could perform the FD task fairly well, that accuracy varied across lemmas (and did not depend only on the number of candidate frames per lemma), and that in a few cases, workers strongly (and correctly) disagreed with gold data. These studies suggest that crowdsourcing for FD is feasible at least on a small scale (about 6 lemmas with a maximum of 5 candidate frames per lemma). The current study adopts and extends many components of that framework to support larger-scale validation of the approach.

2.3.2 Crowdsourcing for WSD

Despite the optimism expressed in Snow et al. (2008) (which included a limited WSD task) and the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (Callison-Burch and Dredze, 2010), relatively few large-scale studies have investigated crowdsourcing for WSD. An important exception is Kapelner et al. (2012), who paid workers to disambiguate 1,000 instances of 89 ambiguous lemmas using the OntoNotes senses (Pradhan et al., 2007), which are relatively

coarse. They found that (1) rephrasing the sense definition improved accuracy, (2) more frequent words were resolved less accurately, and (3) annotators who spent more time per item were less accurate. They also found that all the workers were roughly equal in ability, and those who answered more items did not get more accurate, i.e. there was no measurable practice effect, contrary to the findings of Chen and Dolan (2011), who paid more for better work and tried to retain the more accurate workers.

2.4 Other crowdsourcing for semantics

Few precedents exist for crowdsourcing complex semantic tasks. Bernstein et al. (2010) describe Soylent, a word processor that uses workers on AMT to help writers improve their text. They used a find-fix-verify pattern to iteratively evaluate and refine the quality of tasks like text paraphrasing and summarizing. DuoLingo (von Ahn, 2013) turns translation into an educational game, and translates web content using its language learners.

Freebase is a large human curated collaborative knowledge base (Bollacker et al., 2008) of structured data. The schema for Freebase includes types and relationships that are human curated and validated via large scale crowdsourcing (Kochhar et al., 2010). A key methodological finding from this work was to focus on reproducibility as a key criteria when collecting semantic judgments from human annotators (Paritosh, 2012).

3 Supervised crowdsourcing

The findings discussed above provided promising ways of accommodating some challenges of the FD task. Our goals in extending the FD crowdsourcing framework were twofold: (1) adapt previous efforts to accommodate larger-scale annotation; and (2) incorporate multiple kinds of supervision, broadly construed. We discuss each of these below.

3.1 Scaling up frame disambiguation

We adopted the basic frame-sorting paradigm of Hong and Baker (2011), organizing tasks around specific lemmas. In each task, a set of sentences (each including the target lemma) was presented along with a set of candidate frames (each known to be associated with the target lemma).

Several challenges arose in expanding from these small-scale experiments to less constrained conditions: The 32 lemmas used for our pilot study typically had 3-4 candidate frames but in some cases as many as 10, necessitating an interface that could flexibly accommodate the need for detailed frame definitions within a limited space—while trying to avoid sensory overload that would likely detract from performance. Figure 1 shows a screenshot of the task user interface.

Another problem came from the need to adapt a resource designed for experts for use in a non-expert context. The prose used in FrameNet frame definitions varies considerably in the degree of technical jargon employed—perhaps as much as annotators varied in their appreciation or effective use of those definitions. Hong and Baker (2011) found improved performance with replacing just the frame name with a more easily interpretable title.

Given the impracticality of abridging the frame definitions for each task, we chose to show them unchanged, but to also provide more example uses and related words for each frame to de-emphasize the technical definitions. (We also explicitly warned annotators about the technical jargon and directed them to focus on example uses.)

Finally, we anticipated that a broader range of lemmas would make the task more difficult in various ways. The potential for more candidate frames per lemma raises the chance of ambiguity and similarity among frames. It also seemed likely that there might be cases that fit none of the presented candidate frames for a lemma, either because the appropriate frame had not yet been created or because the lemma in question had not yet associated with that frame. We thus included extra choices corresponding to these failure modes (”None of the above” and ”I can’t decide”), as well as a way for workers to indicate uncertainty or provide additional comments.

As a general principle we also tried to design the simplest interface and instruction materials possible given the nature of the task and the other constraints above. The final guidelines, defining semantic frames for non-experts and introducing them to the task and UI, are 4 pages—longer than a typical crowdsourcing task, but much shorter than materials for expert annotation. These focus on mechanical aspects of the UI and keep terminology and defini-

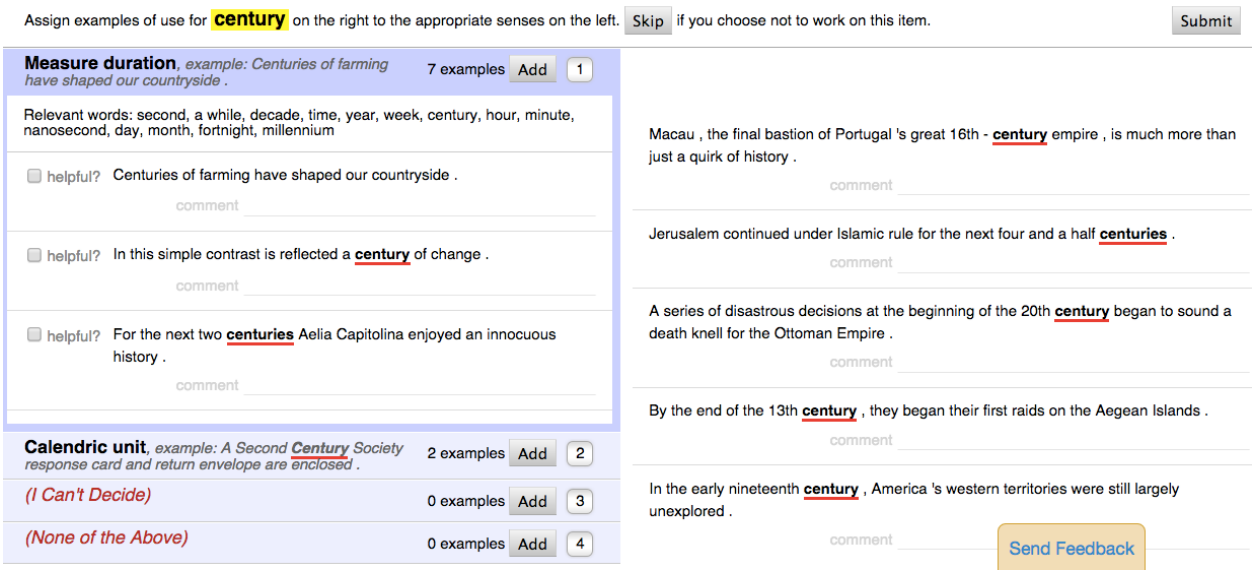

Figure 1: Frame identification task interface for the lemma *century*. Candidate frames (here, **Measure duration** and **Calendric unit**) are shown on the left, each featuring typical examples of usage with the target lemma. The frame definition (not shown in figure) as well as other related words are also available. The examples to be classified are on the right side of the screen.

tions to a minimum.

3.2 Incorporating supervision

In moving to the middle ground of task complexity, we made two broad assumptions that informed how supervision could be introduced.

First, we assumed that the task was complex enough to need some training time, and that annotators with practice and experience would perform better. We thus required a crowdsourcing platform that would allow us to main a relatively stable annotator pool. In contrast to crowdsourcing platforms based on an open marketplace—where anyone is potentially eligible for any task, and no continuity across tasks or workers is guaranteed—we made use of a platform that tracks individual annotators' history and allows some form of communication between task designers and annotators.

This interactive potential of our platform was crucial to our iterative design process: at every stage we were able to conduct small pilot studies that yielded useful qualitative feedback. More broadly, the fact that the same annotators would be working on multiple tasks allowed us to expect and plan for improved performance over exposure to the task—which in turn made it more worthwhile (for both the designers of the task and the annotators) to invest in some amount of training.

Second, we assumed that some gold data would be available for our task. (In our case, it was easy to draw this from the available FrameNet data.) Gold data allows us to follow both conventional wisdom (that people learn best by example) and common practice in (supervised) machine learning of providing explicit training examples of the task being learned. (We have relaxed this assumption in subsequent experiments.)

We use gold data in both *exemplar* and *real-time feedback* form. We lead by (and with) example, by prominently featuring several sentences illustrating each candidate frame. The task UI also allows a mode in which annotators are given explicit positive or negative feedback (in the form of happy or sad faces) indicating whether their frame choice matches the gold data; annotators are allowed to change their frame selection as many times as they would like to. Crucially, we discovered (as in previous work) that gold data occasionally included mistakes, or was potentially ambiguous or uncertain. We thus included explicit means for annotators to indicate disagreement with the apparent gold data (as shown in Figure 3.2), an option that turned out to be quite useful.

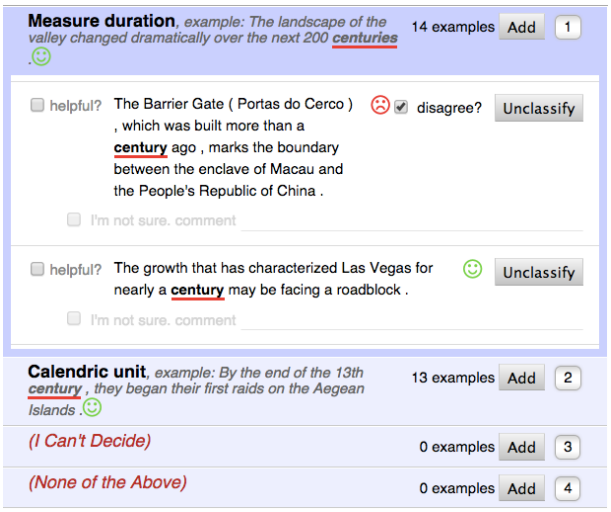

Figure 2: Close-up of task UI used with feedback. Green smiling and red frowning icons indicate correctness of an annotator's selection with respect to the correct (gold) answer, but annotators are allowed to indicate disagreement with the feedback.

4 Frame disambiguation experiment

To investigate how frame disambiguation can be accomplished at scale and with feedback, we used the frame-sorting design and UI described above in several annotation experiments. Below we describe the basic experimental set-up and methodology, followed by our evaluation metrics and results.

4.1 Methodology

We chose lemmas from existing gold examples from FrameNet's full-text annotations, further restricting ourselves to examples from the American National Corpus. We chose 32 target lemmas (occurring in a total of 881 sentences) which satisfy the following conditions:

- At least 15 occurrences in the corpus.
- More than 1 candidate frame for each lemma. The actual number of candidate frames per lemma ranged from 2 to 10 (average 3-4).
- At least 3 examples of the lemma's use in each candidate frame.

The first restriction above (15+ occurrences) was made purely to create tasks of a reasonable size for evaluation; tasks with significantly fewer occurrences have been run with no effect on results.

The second restriction was intended to focus the task on *disambiguation* among multiple frames rather than simply *validation* of a single frame (though other experiments included validation cases). Note that of the current 10K lemmas in FrameNet, 1900 (19%) are polysemous (i.e., associated with more than one frame). These lemmas are thus relatively more ambiguous than the average lemma in FrameNet.

The final restriction, on the number of exemplars available to be shown for the task, was made to facilitate the testing of the feedback condition. Note that more general versions of the task could be run with fewer (or even no) exemplars, or expert annotators could supply those needed.

4.1.1 Experimental design

We used a 2x2 within-subjects factorial design. The lemmas were randomly split into two equal batches (n=16): *No Feedback* and *Feedback*. In the *Feedback* condition, the annotators received real-time positive or negative feedback in response to their sorting actions, based on whether their action matched the gold answer, while no such feedback was provided for lemmas in the other condition. Each annotator performed the task for each lemma, and each lemma was presented with the same type of feedback to all annotators. Each lemma was presented to at least 7 annotators. In both conditions, the annotators were allowed to undo and change their sorting, and every annotator action was logged.

The annotators were randomly allocated to two equal-sized groups: Group 1 and Group 2. Annotators from Group 1 were presented the *Feedback* batch of exemplars before the *No Feedback* batch; and annotators from Group 2 were presented *No Feedback* before the *Feedback* batch. This gives us fully counterbalanced, within-subjects data for comparison of performance across conditions.

4.2 Analysis

We focused our analyses on how **accuracy**—that is, correctness with respect to gold data—varied based on two factors:

Feedback. This is the main dimension we varied across experimental conditions. We compare the difference in performance acrosss *Feedback* and *No Feedback* conditions. We further distinguish the

Feedback condition into two subcategories: Since the task UI allowed annotators to change their selection (potentially in response to gold feedback), we were able to record each frame choice and thus track how well annotators in the *Feedback* condition performed on their first choice for a given item (which we call the *Pre Feedback* condition), as well as what they eventually settled upon (which we call the *Post Feedback* condition).

Number of annotators. We also compared accuracy across different numbers of annotators, ranging from 1 to 7 annotators.

We measured accuracy of the chosen frame against the gold-annotated frame. Our resolution policy was to require a threshold of 75% inter-rater agreement as the minimum for which a resolved answer would be considered usable.

4.3 Results

Figure 3 shows the mean accuracy for the three possible feedback conditions, and Figure 4 shows precision results for different numbers of annotators per lemma (n=1 to 7).

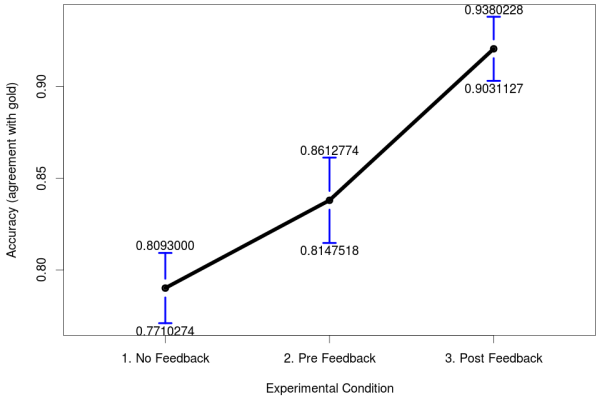

Figure 3: Mean annotator accuracy across three experimental conditions: (1) *No Feedback*, for annotators who received no feedback. (2) *Pre-Feedback*, the accuracy of annotators' first response **prior to** receiving feedback based on gold data. (3) *Post-Feedback*, the accuracy of annotators' final response after receiving feedback, and after any number of revisions. Note that the Post-Feedback accuracy is significantly less than 1.0, showing that annotators have developed strong enough opinions to disagree while learning via the same gold data.

Figures 5 and 6 show individual annotator responses for two lemmas, *like* and *century*. These were both typical in exhibiting a fairly clean division of responses between the candidate frames: i.e., the usages were straightforward to disambiguate. The latter example also includes a panel displaying individual responses, including annotator's disagreement with feedback and frame selection history.

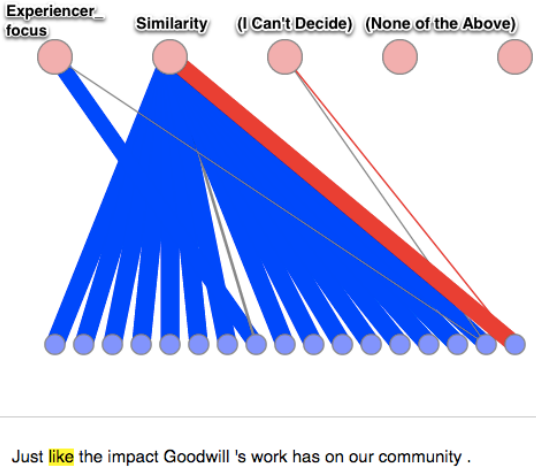

Figure 5: Results for the lemma *like*. The nodes in the top row correspond to candidate frames (Experiencer_focus and Similarity) and three problem conditions("I can't Decide", "None of the above", and an unmarked "Other"). The nodes in the bottom row correspond to classified sentences; lines between nodes in the top and bottom rows represent annotator choices, with thicker lines corresponding to more annotators making that choice. This situation was typical: most sentences had a strong majority for one of the two expected frames, with a few outliers expressing indecision or otherwise disagreeing with the crowd. The red line highlights the results for the single sentence shown below.

We discuss our findings below: Findings 1-3 concerning the effect of feedback, and Finding 4 concerning the effect of number of annotators.

Finding 1. Feedback improves annotator accuracy. Unsurprisingly, we found that feedback improved accuracy: the mean annotator accuracy in the *No Feedback* condition was 0.78, *Pre Feedback condition* was 0.81, and *Post Feedback* condition was 0.92. All differences are significant ($p < 0.0001$). Figure 3 shows the differences between means across the three conditions. In addition, feedback decreased variance in annotator behavior significantly, i.e., the annotators had converged to more

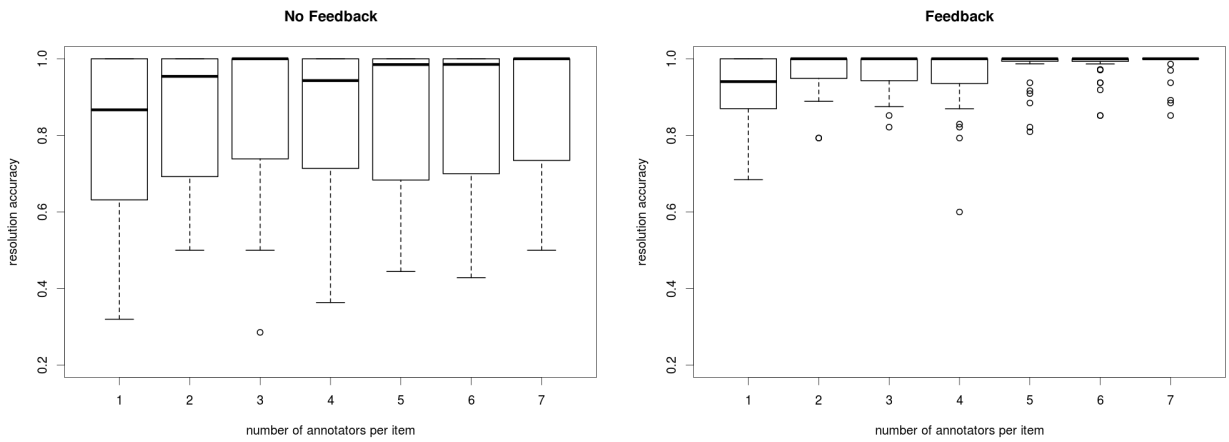

Figure 4: Accuracy of resolutions by number of annotators, in the *No Feedback* (left) and *Feedback* (right) conditions. Box and whisker plots show median (marked by a heavy bar) and variance (indicated by box size) of accuracy across all lemmas. The resolutions are computed by combining independent answers from multiple annotators using a plurality threshold of 0.75.

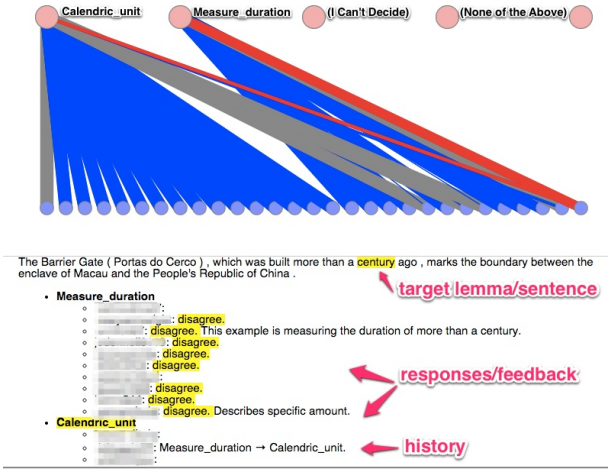

Figure 6: Results for a task with lemma *century*. The responses for individual annotators (names masked) are displayed below, showing that many explicitly disagreed with the gold feedback (some providing additional justification). Note that the history of choices made is shown in one case, also suggesting some uncertainty. This example was one of several that further investigation revealed to be an error in the gold data.

reliable performance. Figure 4 shows two box and whiskers plots of resolution accuracy by number of annotators. There is much wider variance in annotator behavior in the No Feedback condition, as indicated by longer boxes and whiskers.

Finding 2. Feedback works even with imperfect gold data, and can be reliably used to correct it. Our crowdsourced resolutions were significantly *better* than the gold data that was used to train the annotators. In all conditions, annotators were allowed to change their responses; thus, those in the feedback conditions could in theory have performed at 100% accuracy by adhering strictly to the feedback. We were surprised to find, however, that the average accuracy even with feedback was less than perfect—an indication that annotators sometimes chose not to adhere to gold data. We were aware that there might be some errors in the gold data, and allowed and encouraged the annotators to disagree with the feedback.

To investigate cases in which the annotators reliably disagreed with the gold, we asked experts to manually validate gold data for sentences with a resolved answer from the crowd, which was 385 sentences (87.10%). (Recall that we required agreement of 0.75 to be considered resolved.)

Table 1 shows the proportions of validated accuracy of resolved judgments. We found that in most cases (93.77%), the crowd (correctly) agreed with gold. But in some cases (4.94%), the crowd disagreed with gold that turned out to be incorrect. In other words, the crowd was nearly always vindicated when they strongly agreed that the gold was incorrect—and they were overall correct 98.70% of the time.

	number	percent
Correct resolution, valid gold	361	93.77
Correct resolution, invalid gold	19	4.94
Incorrect resolution, valid gold	2	0.52
Incorrect resolution, invalid gold	3	0.78

Table 1: Accuracy of resolved judgments (total 385) based on validated gold data. The top two lines reflect all cases in which the crowd was correct, either in agreement or disagreement with gold data. The bottom two lines reflect very rare cases of incorrect crowd resolutions.

This finding suggests that a richer framework can support crowdsourced semantic annotations even with imperfect data; even better, reliable crowdsourced signals might be an effective avenue to the discovery and correction of imperfect gold data.

Finding 3. Even first responses improve with feedback. Figure 3 shows that the *Pre Feedback* condition was significantly better than the *No Feedback* condition: that is, there seemed to be a boost to performance even on annotator's first guesses (before receiving any feedback). This result suggests that feedback may have had effects that spread beyond the current item, such that subsequent items were learned faster. One possible explanation for this apparent learning based on prior feedback is that there may be increased attention due to the expectation of feedback, such that the annotator homed in more quickly on the correct concept. These hypotheses need further examination.

Finding 4. More annotators produce better results. Unsurprisingly, more is better: resolution accuracy increases with the number of annotators in all conditions. The mean resolution accuracy is higher in the *Feedback* condition, which is as expected since per-annotator accuracy is higher in that condition. In fact, performance was fairly high (in both conditions) with as few as three annotators, but variance in resolution accuracy was significantly lower in the *Feedback* condition, further establishing the effectiveness of feedback. This difference is important, since both mean and variance affect crowdsourcing cost in terms of redundancy required.

5 Discussion and future directions

Our challenge was to devise effective and scalable ways of training annotators to perform the relatively complex task of frame disambiguation. In this paper we have leveraged insights about human learning, in particular the value of exemplars and feedback (early, often and even imperfect), to create a novel crowdsourcing approach suitable for more complex tasks. A key feature of this approach is that it emphasizes examples over explicit instructions, tapping into the cognitive capacity to learn deeply from a limited amount of data. It further exploits supervision, particularly in the form of real-time feedback.

We demonstrated that real-time feedback can substantially increase mean annotator accuracy and dramatically increase inter-annotator agreement. Our experiments also showed the surprising result that even feedback based on imperfect gold data is effective for training annotators—and that they can learn to produce resolutions of higher accuracy than the gold data they trained on. This suggests that we can train annotators with tarnished gold, and as part of that process even improve the gold data.

Besides being valuable in its own right as a version of word sense disambiguation, this task is also a small step on the road to full frame semantic annotation. We are currently piloting the task for the next step toward full frame annotation (frame element annotation), applying the same principles of feedback and supervision.

More generally, the supervised crowdsourcing paradigm developed here explores a useful middle ground of expertise, one we believe to be suitable for many semantic annotation tasks too complex for standard transient crowdsourcing. An effective way of producing such data on a large scale using faster, less expensive methods has great potential for easing the semantic bottleneck and facilitating progress toward richer natural language understanding.

Acknowledgments

We gratefully acknowledge support from Google in the form of a Google Faculty Research Fellowship to Collin Baker. On the FrameNet team, we thank Michael Ellsworth for insights on the annotation process and gold data validation, and Warren McQuinn for gold data validation. At Google, we thank Binbin Ruan and Xiaoming Wang for their help on the UI, and Dipanjan Das, Michael Tseng, Russell Lee-Goldman, Ed Chi, Jamie Taylor, Eric Altendorf,

John Giannandrea and Amar Subramanya for useful discussion and feedback.

Thanks also to the reviewers for very thoughtful, constructive comments. Any opinions or errors are those of the authors alone.

References

Michael S. Bernstein, Greg 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 the 23nd annual ACM symposium on User interface software and technology*, pages 313–322. ACM.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Chris Callison-Burch and Mark Dredze, editors. 2010. *Proceedings of the NAACL/HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA, June. ACL.

David L. Chen and William B. Dolan. 2011. Building a Persistent Workforce on Mechanical Turk for Multilingual Data Collection. In *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*, August.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic Frame-Semantic Parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference*, Los Angeles, June.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2013. Frame-Semantic Parsing. *Computational Linguistics*, 40(1).

Charles J. Fillmore and Collin F. Baker. 2010. A Frames Approach to Semantic Analysis. In Bernd Heine and Heiko Narrog, editors, *Oxford Handbook of Linguistic Analysis*, pages 313–341. OUP.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the Crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 742–747, Sofia, Bulgaria, August. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Jisup Hong and Collin F. Baker. 2011. How Good is the Crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 30–37, Portland, OR, June. ACL.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-based Semantic Role Labeler for Swedish. In *Proceedings of Coling/ACL 2006*, Sydney, Australia, July 17-21.

Adam Kapelner, Krishna Kaliannan, H. Andrew Schwartz, Lyle Ungar, and Dean Foster. 2012. New Insights from Coarse Word Sense Disambiguation in the Crowd. In *Proceedings of COLING 2012: Posters*, pages 539–548, Mumbai, India, December. COLING.

Shailesh Kochhar, Stefano Mazzocchi, and Praveen Paritosh. 2010. The anatomy of a large-scale human computation engine. In *Proceedings of the acm sigkdd workshop on human computation*, pages 10–17. ACM.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, March.

Praveen Paritosh. 2012. Human computation must be reproducible. In *CrowdSearch*, pages 20–25.

Miriam R. L. Petruck and Gerard de Melo, editors. 2014. *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, Baltimore, MD, USA, June. Association for Computational Linguistics.

Sameer Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. *International Journal of Semantic Computing*, 1(4):405–419.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA. Distributed with the FrameNet data.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 EMNLP*, pages 254–263, Honolulu, HI, October. ACL.

Luis von Ahn. 2013. Duolingo: Learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, pages 1–2. ACM.