Crowdsourcing Breakthroughs for Language Technology Applications
AAAI Technical Report WS-15-24

Linguistic Wisdom from the Crowd

Nancy Chang, Russell Lee-Goldman, Michael Tseng[*]
Google Inc.
1600 Amphitheatre Parkway
zMountain View, CA 94043
{ncchang, rlg, michaeltseng}@google.com

Abstract

Crowdsourcing for linguistic data typically aims to replicate expert annotations using simplified tasks. But an alternative goal—one that is especially relevant for research in the domains of language meaning and use—is to tap into people's rich experience as everyday users of language. Research in these areas has the potential to tell us a great deal about how language works, but designing annotation frameworks for crowdsourcing of this kind poses special challenges. In this paper we define and exemplify two approaches to linguistic data collection corresponding to these differing goals (*model-driven* and *user-driven*) and discuss some hybrid cases in which they overlap. We also describe some design principles and resolution techniques helpful for eliciting linguistic wisdom from the crowd.

1. Introduction

Language technologies have come a long way: speech and language interfaces designed to support naturalistic interactions are increasingly common across a wide variety of platforms and applications. While the degree to which such technologies succeed in mimicking human linguistic behavior remains in the eye (and ear) of the interlocutor, systems that work even occasionally, in limited domains, are both intuitively appealing and suggestive of their potential for improvement.

The key force driving this progress has been the increasing availability of large-scale data resources, coupled with advances in statistical machine learning. Systems performing speech recognition, machine translation and syntactic parsing, for example, have relied on training data available from resources like the Wall Street Journal, Europarl and the Penn TreeBank, respectively. Over the last decade, progress on more semantically oriented tasks (e.g., question answering and semantic role labeling) likewise point to the need for annotated data of sufficient quality and quantity to train statistically learned models.

But obtaining data on this scale can be both time-consuming and expensive, especially since linguistic annotation has traditionally been the domain of trained experts. Many tasks require a lengthy iterative process to define annotation standards suitable for a specific language phenomenon and then train contributors to produce data. These challenges are exacerbated by longstanding theoretical divisions that make it difficult to come to a consensus about how to represent aspects of meaning and usage. Moreover, even when the target annotations are clearly defined, the sheer volume of annotations can be a significant obstacle, especially for low-resource languages.

One potential solution to the data bottleneck is to exploit crowdsourcing as an alternative to expert annotation. Across various domains, the use of a large number of non-expert judgments in lieu of a small number of expert judgments has proven a scalable and inexpensive means of amassing large-scale data needed for complex tasks.

Drawing on the fruits of human computation to obtain linguistic annotation data seems like an ideal fit, particularly since annotation frameworks designed for non-experts could extend naturally to other languages and thereby provide the means of tapping into a large pool of competent language users worldwide. To date, however, tasks and techniques addressed in the human computation community tend, for natural reasons, toward those involving relatively simple judgments that require no (or minimal) training and bear little resemblance to the kinds of complex structured annotations that are the goal of expert linguistic annotation.

This paper describes some exploratory investigations of the use of crowdsourcing for linguistic data collection, with the hope of bringing the two relevant communities closer together for their mutual benefit. In particular, the relative complexity of linguistic data collection poses novel challenges to the human computation community, ones we believe are crucial to solve for building systems that strive for complex, human-level behavior. But we also believe that crowdsourcing frameworks and techniques have great potential for fundamentally transforming how we collect linguistic data, and even how we define it.

We first situate our investigations by identifying two basic categories of linguistic crowdsourcing (*model-driven* and *user-driven*). In practice, these categories can overlap; we illustrate several hybrid linguistic crowdsourcing tasks. We then summarize some techniques and strategies we have found useful for tasks of this kind, in both task design and data resolution. Finally, we mention some theoretical con-

[*] Authors contributed equally to this paper and are listed in alphabetical order.

siderations affecting potential goals and strategies suitable for various kinds of linguistic crowdsourcing.

2. Background: two kinds of data collection

Natural language technologies differ in the kinds of data they require, as well as how closely these match the behaviors and intuitions that drive everyday language use. In this section we define two broad categories of data collection:

- **model-driven**: deriving expert annotations from non-expert annotations in a well-understood problem space
- **user-driven**: eliciting the knowledge of non-experts with respect to a novel problem space

These categories correspond to two distinct motivations for collecting linguistic data; they also have different historical antecedents. Below we describe each in more detail.

Model-driven data collection

As in many other domains, crowdsourcing for linguistic data is often oriented toward replicating expert annotation using the more scalable population of non-experts, typically to produce large-scale labeled data for training or testing NLP algorithms. In such cases, the format of the labeled data and standards governing annotation arise from a specific natural language application and presume a particular underlying linguistic model.

We call this *model-driven* data collection, a mode of research driven by the needs of the modeling task at hand. Examples of tasks traditionally performed by experts include part-of-speech tagging, syntactic parsing, named entity recognition and co-reference resolution.

Of course, some of these (e.g., syntactic parsing, part-of-speech tagging) range far afield from everyday language use and thus require significant training (e.g., to train contributors to produce phrase structure trees like those of the Penn TreeBank). Other tasks require less specifically linguistic training, but nonetheless involve annotating tacit knowledge or inferences (co-referring entities, named entities) that are typically not explicit. Training might thus be needed to ensure consistent annotations.

The appeal of model-driven data collection is that the tasks tend to be well-defined and clearly circumscribed, both in terms of the target data and the guidelines previously established through expert annotation. The utility of the data for NLP applications is also typically well-attested. The challenge for these tasks lies in developing annotation frameworks (including user interfaces, guidelines and training processes) that adapt the task appropriately for non-experts, so that their annotations can substitute for or supplement those of experts.

User-driven data collection

A rather different situation applies when the goal is to discover how users (that is, humans) behave. Here, data collection is often a tool for exploring a novel problem space for which models and target representations are not yet well-established. Such is the case for many aspects of linguistic knowledge, most notoriously those involving the shifting sands of meaning and interpretation. Compared with the relatively stable theoretical ground of model-driven data collection, annotations involving semantics and inference tend to involve more variation across speakers, usage contexts and applications.

For such cases, it may be more appropriate to design tasks to elicit insights from a typical language user that reflect natural language use and interpretation. We call this *user-driven* data collection, designed to elicit the knowledge that a typical language user (as represented by a set of non-expert contributors) draws upon in everyday language use, often without being aware of it.

In this mode, careful design remains important to ensure meaningful comparison and analysis of contributors' responses, but the task should also hew as closely as possible to natural usage contexts. The role of the researcher shifts away from simulating "expert" behavior (because there is no such standard in the first place) to encouraging contributors to tap into their intuitions and "commonsense" reasoning in a way that makes it possible to find a signal in all the data collected.

User-driven data collection can be seen as more akin to psycholinguistics experiments than to traditional annotation efforts. This idea resonates with the findings of (Munro et al. 2010), who provide a number of illustrations of how crowdsourcing can go beyond mere data annotation to yield rich information about the nature and distribution of semantic intuitions, at a comparable quality to expert insights.

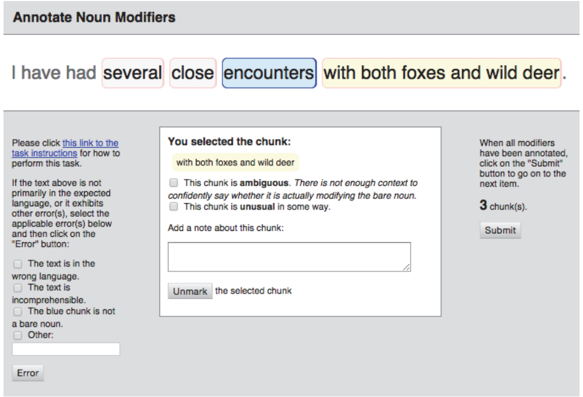
One example of user-driven data collection is a task described by (Barua and Paritosh 2015), designed to elicit "commonsense properties" of question–answer pairs from web corpora such as Yahoo! Answers and StackExchange. They tested a panoply of prompts to probe features of a given question and answer, such as whether a question is seeking a factual answer or opinion, how long the expected answer would be, and the perceived expertise of the answer's writer.

While not a traditional linguistic task, this task relied on non-experts' tacit expectations and inferences using general and commonsense knowledge. Given the lack of an established model and ontology for these properties, the priority was to collect and analyze a large volume of responses to discover properties with high inter-rater reliability and thus inform future refinements to the task.

Note that many long-established NLP tasks can be reinterpreted as examples of tasks best suited to user-driven data collection. Speech recognition and machine translation, for example, have clear analogues in pre-existing functions traditionally performed by humans (e.g., speech transcription and translation)—ones requiring no specific linguistic background (though contributors may have a variable degree of task expertise).

3. Hybrid linguistic tasks

The two approaches to data collection defined above are not mutually exclusive: specific language tasks often incorporate techniques from both. This section describes several tasks we have designed with this hybrid character.



Annotate Noun Modifiers

I have had several close encounters with both foxes and wild deer.

Please click this link to the task instructions for how to perform this task.

If the text above is not primarily in the expected language, or it exhibits other error(s), select the applicable error(s) below and then click on the "Error" button:

□ The text is in the wrong language.
□ The text is incomprehensible.
□ The blue chunk is not a bare noun.
□ Other:

Error

You selected the chunk:

with both foxes and wild deer

□ This chunk is **ambiguous**. *There is not enough context to confidently say whether it is actually modifying the bare noun.*
□ This chunk is **unusual** in some way.

Add a note about this chunk:

Unmark the selected chunk

When all modifiers have been annotated, click on the "Submit" button to go on to the next item.

3 chunk(s).

Submit

Figure 1: The last stage in the noun phrase sequence.

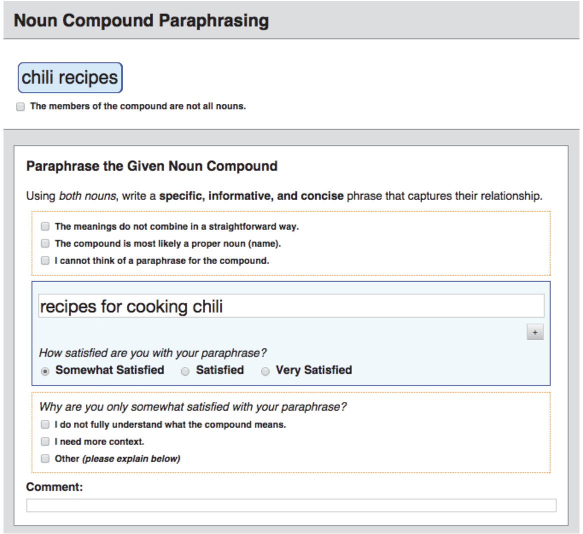
Noun phrases (referring expressions)

One hybrid task aims to tap into non-experts' knowledge corresponding to the identification of **noun phrases** (that is, expressions whose *syntactic head* is a noun). While existing data of this form collected from experts is available for some languages, many lower-resource languages lack sufficient data of this kind to support part-of-speech tagging and syntactic parsing. We were thus interested in developing methods for obtaining such data from non-experts in a cross-linguistically robust way.

We piloted an annotation sequence to collect noun phrases from a corpus of sentences, illustrated in Figure 1 for English (though the annotation UI has been used for other languages). In fact, noun phrases often exhibit complex internal structure (as in the example in the figure), which in traditional grammatical terms can involve determiners (*several*), adjectives (*close*), modifying prepositional phrases (*with both foxes and wild deer*), adjuncts, and various possible relationships among these.

The challenge here was to make this task suitable for non-experts, preferably without having to explain theoretical syntactic concepts. We drew on several strategies:

- Task decomposition. We divided the task into two stages: first, identifying **bare nouns**, corresponding to the syntactic head of the noun phrase (e.g. *encounters* in the figure); and second, identifying the bare nouns' **modifiers**, corresponding to the dependent expressions (i.e., the other highlighted chunks in the figure).
- Accessible terminology. We found it more effective to avoid syntactically oriented terms and appeal instead to the more intuitive semantic notions of reference (what kind of entity is being referred to) and modification (what phrases in the text are describing or modifying that entity).
- Reinforcement training. We provided feedback showing differences between submitted annotations and gold-standard annotations.
- Structural synthesis. The contributors' responses were motivated by semantic or functional criteria and thus did not always align perfectly with standard syntactic conventions (for example, contributors varied in where they placed segment boundaries). Through a custom resolution algorithm, however, we were able to stitch contributors' semantically motivated responses together into the desired composite structure, corresponding to a (syntactic) noun phrase or (semantic) referring expression. (See Section 4 below for more details.)



Noun Compound Paraphrasing

chili recipes

□ The members of the compound are not all nouns.

Paraphrase the Given Noun Compound

Using *both nouns*, write a **specific, informative, and concise** phrase that captures their relationship.

□ The meanings do not combine in a straightforward way.
□ The compound is most likely a proper noun (name).
□ I cannot think of a paraphrase for the compound.

recipes for cooking chili

+

How satisfied are you with your paraphrase?
● Somewhat Satisfied ○ Satisfied ○ Very Satisfied

Why are you only somewhat satisfied with your paraphrase?
□ I do not fully understand what the compound means.
□ I need more context.
□ Other *(please explain below)*

Comment:

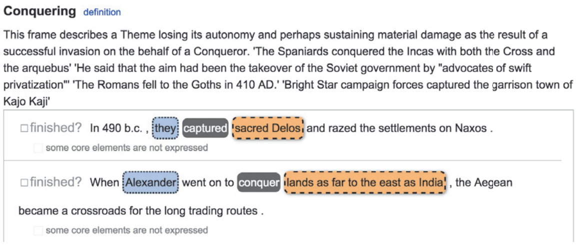
Figure 2: The noun–noun compound paraphrasing task.

The resulting hybrid framework effectively performs model-driven data collection of conventional syntactic structures by applying user-driven data collection techniques that recast the task in more semantically oriented, usage-based terms. It has been applied to collect structured noun phrase annotations for English and other languages.

Noun–noun compounds

The tension between satisfying the requirements of an established model and data format versus encouraging contributors to provide exploratory insights can accrue during the course of a task's development. While piloting a task to explore the latent conceptual relationships in **noun–noun compounds** (a *kitchen table* is a table that *is located in* a kitchen, whereas a *dinner table* is a table that is *used for the purpose of serving* dinner), we initially encouraged contributors to provide paraphrases in any format that was "specific, informative, and concise" (Figure 2).

The paraphrases were intended to train a noun–noun compound classification system (i.e., one that chooses which of a set of candidate relationships holds for a particular noun–noun compound). To encourage consistency in the training data that would facilitate learning, the modeler wished to restrict the paraphrase format to a smaller space of possibilities. Since this restricted format sometimes made for unintuitive paraphrases, we developed more extensive guidelines and "gold" paraphrases that were used as examples for reinforcement training (i.e., given as immediate feedback during the task).



Conquering definition

This frame describes a Theme losing its autonomy and perhaps sustaining material damage as the result of a successful invasion on the behalf of a Conqueror. 'The Spaniards conquered the Incas with both the Cross and the arquebus' 'He said that the aim had been the takeover of the Soviet government by "advocates of swift privatization"' 'The Romans fell to the Goths in 410 AD.' 'Bright Star campaign forces captured the garrison town of Kajo Kaji'

□finished? In 490 b.c. , they captured sacred Delos and razed the settlements on Naxos .

□ some core elements are not expressed

□finished? When Alexander went on to conquer lands as far to the east as India , the Aegean became a crossroads for the long trading routes .

□ some core elements are not expressed

Figure 3: Example annotations from the frame element annotation task, for the Conquering frame. Text highlighted in blue marks the conqueror, while text in orange marks the conquered party.

As the scope of the task shifted from user- to model-driven, the techniques we used shifted from channeling commonsense understanding of conceptual relationships toward training contributors to develop operational models more closely aligned with the desired format.

Frame semantics

A third area for which we have developed hybrid annotation tasks is that of frame semantics, of the form pioneered by the FrameNet project (Fillmore and Baker 2010). This domain seemed particularly ripe for crowdsourcing, since the underlying intuitions reflect domain-general, non-expert knowledge of the meanings of words and sentences. Most frame annotation to date, however, has been produced by trained experts.

As in the other cases, the challenge was to adapt frame-based tasks into ones suitable for non-experts (see (Chang et al. 2015)). The case of **frame disambiguation** (choosing which of a set of candidate semantic frames is the relevant sense of a target lexical item used in a sentence) draws heavily on providing real-time feedback during training and emphasizing examples over instructions. In ongoing work on **frame element annotation** (identifying the textual segments associated with different semantic roles, illustrated in Figure 3), we are pursuing similar techniques, in combination with exploiting better resolution techniques to robustly identify the desired segments.

Both annotation tasks are hybrid in the sense explored here. They are model-driven in that they match the representational format established by trained experts and are suitable for machine-learned models (i.e., semantic role labeling systems as pioneered by (Gildea and Jurafsky 2002)). Moreover, the frame-based vocabulary used to identify these relations is drawn from a taxonomy designed for experts (and exemplified by the frame definition for Conquering shown in the figure). But the semantic and conceptual relations that are the focus of frame-semantic annotation are rooted in everyday knowledge accessible to typical speakers, not just experts; these thus qualify as user-driven.

It would be most accurate to say that the information targeted in this task is non-expert (hence more user-driven), but that the form in which it is expressed is expert (using a specialized vocabulary for both frames and frame elements; choosing frame element boundaries following specialist guidelines). Previous work by (Hong and Baker 2011) had demonstrated that simplifying the jargon (by employing more user-friendly frame names) improves performance in frame-based crowdsourcing experiments. For our purposes, however, we decided that translating the entire vocabulary would not be practical. We thus sought other ways to simplify the task for non-experts while also satisfying the model-based need for data of a particular format. (See Section 4 below for more details.)

4. Techniques for linguistic crowdsourcing

As illustrated by the sampling of tasks above, we have found a variety of techniques useful for collecting linguistic data using human computation frameworks, across the model-driven, user-driven and hybrid styles of collection.

These techniques can be divided into ones that improve **task design**, exploiting insights about the human learning capacity to achieve robust and high-quality results; and ones that improve **resolution**, which compensate for the relatively greater potential for inter-annotator variance or error in the work of non-expert contributors.

Task design techniques

We have found that general principles of design, learning and supervision are useful aids in guiding the creation and refinement of crowdsourced tasks. Regardless of the overall task orientation (i.e., model- vs. user-driven), simpler tasks that minimize cognitive load, context-switching and lag time between actions and feedback tend to reduce the chance for errors and confusion.

Examples of techniques that can be straightforwardly extended to the linguistic domain include the use of example-based training and real-time feedback (to allow reinforcement learning), as in the **supervised crowdsourcing** framework described in (Chang et al. 2015); and **task decomposition** (decomposing a cognitively complex task into a sequence of simpler operations such as filtering or ranking), as in (Callison-Burch and Dredze 2010).

For model-driven data collection, the emphasis is on developing an environment in which non-experts effectively behave (at least collectively) like experts. To that end, these techniques serve to create a *process* that guides non-experts toward developing implicit mental models that yield expert-level annotations. These approaches rely on non-experts' general cognitive abilities, thus sidestepping the need to train them toward explicit linguistic annotations.

For user-driven data collection, the emphasis is on developing an environment in which non-experts (who can be thought of as experts at everyday language use) can behave as naturally as possible.

Below we elaborate on several principles of task design that we have found to improve result quality.

Simplify user interfaces. All else being equal, a simpler task UI is better. The less one has to think about the elements of the UI, the more one can focus on the task at hand. We believe this is especially true when the aim is to elicit natural reactions and judgments about language. Nothing in the UI should get in the way of tapping into contributors' intuitions in textual interpretation.

The best way to put this principle into practice, we have found, is simply to ensure sufficient usability testing and iteration—that is, test early and often. Though it may seem obvious, it is crucial to involve all parties—including the clients requesting linguistic data (whether to serve a particular model-driven need or for user-driven data collection), the task designers, expert annotators if available (particularly ones with relevant linguistic knowledge) and representative (non-expert) contributors—in piloting enough examples of a task early enough to uncover and address UI problems. As a bonus, such problems often reveal subtler issues related to underspecified task definitions, difficult boundary cases and other unanticipated aspects of the problem space.

Emphasize examples over guidelines. The written instructions for a task, whether presented in a separate document or displayed within the task UI, represent another key method for guiding contributors' work. We have found, however, that detailed guidelines and definitions are not always the most appropriate method for presenting a task.

Detailed definitions and instructions are most suitable for model-driven tasks, particularly in cases where the desired annotation must conform to a specific format (i.e., one that is equivalent to or easily transformed into the expert annotations expected by the model). By contrast, relatively simpler guidelines with a minimum of jargon are better suited to the user-driven approach.

But in both cases, the most effective form of guidance tends to be the inclusion of plentiful examples or sample responses, either alongside the guidelines or within the task itself. Examples provide a more readily understood means of delimiting and clarifying the task than definitions. They are especially informative for linguistic tasks, since contributors can build on their extensive everyday experience with language to internalize the parameters of the task.

One consequence of emphasizing examples over guidelines is that contributors may have somewhat less uniform or consistent responses. While such an outcome is generally undesirable for model-driven data collection, it may in fact be an advantage when it comes to user-driven data collection. When contributors have the freedom to interact with the task items on their own terms and derive idiosyncratic ways of modeling the space of judgments, the resulting data exhibits a richness that reflects how people use and understand language in more natural, everyday contexts. The opportunity to discover the true patterns underlying language behavior is ultimately one of the great boons of crowdsourcing in this domain.

Feedback matters. One of the most effective techniques for improving both the quality of contributor responses and the speed with which they reach competence on relatively complex and challenging tasks is to provide feedback of various kinds—and again, early and often is the key. We have found that contributors appreciate feedback regarding their judgments and performance, which can also help keep them engaged in long and sometimes challenging tasks.

We have developed multiple mechanisms for displaying feedback, especially in or near real-time (as contributors make decisions, or upon submitting individual task items). This technique is limited to tasks where gold annotations are available, either taken from an external corpus (e.g. FrameNet and OntoNotes) or specific to this task (created by the researchers, solicited from other experts or derived from resolving contributors' responses in previous iterations of the task).

Task items associated with gold annotations can then be interspersed with other task items. When such a task item is completed (e.g., a judgment made or a classification completed), contributors immediately see whether their action agrees with the gold. Over time, contributors become more familiar with the distinctions targeted by the task and more confident about their own judgments.

This confidence can be put to especially good use when contributors are given the means to disagree with gold annotations, for two reasons: (1) gold data can contain a surprising number of errors; and (2) the nature of the domain sometimes affords more than one plausible response (as noted above). In more than one case, the signal provided by allowing contributors to have a mind of their own, so to speak, has led to the re-evaluation of our own data and annotation policies, as well as to the revision of gold data from lexical resources.

The frame disambiguation task mentioned above serves as a good illustration of the multiple benefits of using feedback (described in more detail in (Chang et al. 2015)). Incorporating feedback produced higher classification accuracy overall and lower variance across contributors achieved with fewer contributors. But more strikingly, contributors allowed to disagree with gold data proved both willing and able to do so: in the vast majority (86%) of cases in which the contributors' resolved response disagreed with gold data, they had correctly identified incorrect gold data. This use of feedback aligns well with model-driven data collection, and shows how under the right conditions non-experts can learn to match and even improve upon expert-annotated data.

On a more qualitative level, the task allowed contributors to flag several other possible failure conditions (e.g., the inclusion of "none of the above" and "more than one of the above" options to indicate when the given classification options were incomplete or not mutually exclusive). This use of feedback is especially compatible with the user-driven goal of discovering how competent language users understand text.

Finally, feedback can go both ways: we solicit free-form feedback, within the task UI itself as well as "offline" (e.g, via email), as a way to discover the good and bad points of our task design, as well as to cultivate contributors' longer-term interest in and engagement with the task. Again, these suggestions have often revealed hidden assumptions in our tasks or borderline cases that require more careful treatment.

Resolution techniques

In contrast to the task design techniques just described, which focus on how to elicit the desired contributor input, resolution techniques focus on how to aggregate, combine and otherwise resolve these judgments into a summary judgment, i.e., the desired task output.

Many types of linguistic annotation involve tasks suitable for (or adapted to) standard crowdsourcing techniques. For example, linguistic classification tasks can be treated much as any other classification task, so we can similarly aggregate/resolve multiple judgments to achieve higher-quality and more robust results.

The linguistic domain poses some additional challenges (or opportunities) due to the presence of structure, at multiple levels of granularity (e.g. words, phrases, sentences) and across multiple domains (e.g. syntax and semantics). Below we discuss a few of the resolution techniques most relevant to the kinds of hybrid linguistic crowdsourcing tasks we have described.

More is better. The usual advantages of crowdsourcing apply: a large number of contributor judgments can guard against the inherent risk of human error with even the most expert of contributors.

In the context of model-driven tasks, a greater number of judgments warrants greater confidence that the resolved result correctly triangulates to the desired result (perhaps even in disagreement with gold data, as mentioned earlier). Likewise, for user-driven tasks, even those without a predefined correct result, more data can smooth out some of the noise and variability across contributors, as well as provide a richer picture of the range of possible responses.

In cases where the judgment's answer space is scalar or enumerable, resolution typically involves taking a mean or plurality vote; with sufficient contributors, additional statistical information can be exploited (for example, items whose judgments exhibit high variance can be thrown out as irresolvable or escalated to experts). When working with a small contributor pool, it can help to integrate debiasing into resolution by weighting contributors' votes on the basis of their performance on a set of task items with gold annotations, as in (Snow et al. 2008).

Structural synthesis. Many linguistic annotation tasks involve identifying structures defined with respect to a given textual segment. These structures often include internal structure that makes it difficult to apply standard resolution techniques.

For example, the noun phrase task described as one of our hybrid linguistic tasks in Section 3 includes a stage in which contributors select all text spans (or chunks) in a given sentence that they judge to be modifying the given noun. (In the example task item shown in Figure 1, the modifier chunks are *several*, *close* and *with both foxes and wild deer*.) The resulting set of chunks (each a subspan of the given sentence) must still be resolved—and combined, in this case—to encompass the maximal extent of the noun phrase with the given head noun. (In the example, the resolved full noun phrase would be *several close encounters with both foxes and wild deer*.)

As mentioned earlier, the task decomposition strategy taken here means that contributors are never asked to explicitly label the entire complex structure. Rather, we synthesize this structure based on the intermediate results of the two subtasks (which identified bare nouns and modifier chunks, respectively), using the resolution algorithm informally described here:

- For each head noun, filter all identified modifier chunks to preserve only those marked by a (configurable) plurality of contributors.
- Check whether each remaining modifier chunk contains a head noun from another task item; if yes, merge the contained head noun's (remaining) modifier chunks into the modifier chunk that contains it.
- Flatten each head noun and its merged modifier chunks into a single noun phrase.

In a way, the first stage of the resolution is simply a tallying of "votes" on modifier chunks, even though contributors do not explicitly choose from scalar or enumerable values; the second and third stages use the collective evidence of these contributors' votes to synthesize a composite structure that respects a particular linguistic model of phrase structure.

Cross-domain task decomposition. A recurring theme across many tasks we have discussed is the fine line between formal, syntactic structure and usage-based, semantic structure. The close relationship between these different domains of linguistic structure can sometimes be exploited to facilitate complex resolution cases, particularly for hybrid tasks that mix model-driven and user-driven components.

The frame element annotation task mentioned above provides an illustrative example. This task is primarily semantic, since it requires an understanding of the participants, objects and semantic relationships depicted in the sentence—that is, the corresponding structured scene. To demonstrate this understanding, the annotator must label the sentence with respect to a particular frame. In the example below, the three underlined segments correspond to the frame elements (identified below each line) defined in the *Giving* frame:

They	**gave**	the child	several boxes of treats.
giver		*recipient*	*given-object*

This task divides naturally into two distinct (though interrelated) subtasks:

- **chunking**: identifying each textual chunk that corresponds to a frame element (i.e., the underlining above)
- **classification**: categorizing each identified chunk as instantiating a specific frame element (i.e., adding the italicized labels above)

With respect to resolving multiple judgments, the classification subtask requires only standard resolution techniques (taking majority or plurality over judgments). The chunking subtask, however, requires the identification of multiple structural elements; each of these, while semantically motivated, requires a judgment about its exact boundaries.

Contributors can in theory (and do in practice) vary considerably in these boundary judgments. In the example sentence, many options would be reasonable choices for the *object-given*, including *boxes*, *treats*, *several boxes* and *boxes of treats*. This, we believe, is a quite intuitive way of thinking about what a sentence means. ("What was given? Treats.") Similarly, for the first sentence in Figure 3, *Delos* seems just as plausible an answer as *sacred Delos*, and the head noun *lands* would make sense for the second.

Indeed, from a user-driven perspective, all of these answers are conceptually correct. But from a model-driven perspective, one may be considered more correct than the others. In particular, FrameNet gold annotations generally correspond to syntactic constituents (e.g., the full noun phrase for the cases above). Thus, if the goal is to re-create expert frame annotations, contributors would need training (either explicit or implicit, potentially exploiting feedback, examples and other task design techniques) to identify and use syntactic structure to choose the desired boundaries.

When determining the best resolution strategy for a given task, it is thus crucial to consider the goal of the annotation. To support model-driven annotation in this case, for example, we might devise a custom strategy for aggregating judgments into a composite full noun phrase (e.g., take the maximal span covered by the union of chunks).

Happily, this aggregation step bears a strong resemblance to that of the separate noun phrase task described above. We can thus employ a similar technique to stitch together a syntactically preferred whole from multiple (semantically motivated) parts. Even more conveniently, we could simply use judgments from the frame element annotation task (e.g. the word "treats") as input to the noun phrase task.

In sum, decomposing our more complex original task into coherent subtasks with clear goals makes it possible to exploit previously developed tasks and employ different resolution techniques as appropriate. Given the interconnected nature of linguistic structure, we believe that careful task decomposition that is sensitive to differing goals will be crucial for allowing task reuse across a wide array of linguistic phenomena.

5. Discussion

We have found it useful in this overview of linguistic crowdsourcing tasks and techniques to highlight the dichotomy between model-driven and user-driven data collection paradigms. These broadly (but imperfectly) correlate with several other dimensions of variation:

- experts vs. non-experts
- syntactic vs. semantic structure
- definitions and guidelines vs. examples
- training to desired outcome vs. psycholinguistic discovery

These oppositions are rooted in genuine distinctions. Model-driven data collection is geared toward (re)producing a specific format for linguistic information, one traditionally produced by trained experts and with strict formal and syntactic constraints. User-driven data collection tends to target semantic and functional aspects of everyday language use.

But as with many other dichotomies, some of the most interesting possibilities lie somewhere in the middle. The hybrid tasks described above all represent various intermediate cases that integrate multiple motivations and techniques. These yield especially interesting arenas for exploring how different techniques, related to both task design and resolution, might be suitable for different situations.

Perhaps the distinction most readily blurred is that between experts and non-experts, particularly for crowdsourcing linguistic data related to the domains of meaning, context and use. It is specifically in these areas where linguistic structure interacts with everyday experience. Figuring out how to express or extract meanings, negotiating how and when to use or avoid certain expressions and identifying essential parts of an extended passage are "tasks" that people do every day, intuitively and mostly successfully. Unbiased by preconceived notions of what meaning is or how it works, the crowd is not just a useful, but possibly the optimal way to discover the important categories of linguistic analysis in these domains. It is precisely in "the crowd" where language meanings are (co-)defined and used. That is, non-experts *are* experts, at least collectively, in the domain of everyday language use.

As suggested by several of the directions described above, we believe that it is also possible to approach more abstract, formal domains of linguistic knowledge from the more familiar ground of semantics and usage. (The noun phrase task described above represents a small move in this direction.) These domains do not seem especially amenable to crowdsourcing, since they involve linguistic structures that are defined primarily inwardly, with respect to other formal objects—such as the structure of sounds, words and syntactic constituents.

We are optimistic, however, that this kind of structural information can be discovered by tapping into the wisdom of the crowd. Some ingenuity may be needed to devise tasks that exploit (or reveal) the links between such structures and everyday language use, but such explorations would resonate well with some (relatively) recent movements within theoretical linguistics to take seriously the ways in which linguistic structure must be understood in terms of its use and its users. The resulting body of collected linguistic wisdom would represent a significant step toward closing the gap between computational models and the human behaviors they are intended to approximate.

Acknowledgments

We thank David Huynh, Ji Li, Mehdi Manshadi, David Hall and Praveen Paritosh for helpful discussion of our annotation program and contributions to the design of our annotation tasks, along with colleagues in Google Machine Intelligence whose crowdsourcing tools made this work possible. We are also grateful to the crowd of contributors whose annotations and feedback informed the development of the tasks and broader framework described here.

References

Barua, A., and Paritosh, P. 2015. Using commonsense for deeper understanding of complex question answer content. In *WebQA, SIGIR 2015.*

Callison-Burch, C., and Dredze, M., eds. 2010. *Proceedings of the NAACL/HLT 2010 Workshop on Creating Speech* *and Language Data with Amazon's Mechanical Turk*. Los Angeles, CA: ACL.

Chang, N.; Paritosh, P.; Huynh, D.; and Baker, C. F. 2015. Scaling semantic frame annotation. In *Proc. 9th Linguistic Annotation Workshop, NAACL.*

Fillmore, C. J., and Baker, C. F. 2010. A Frames Approach to Semantic Analysis. In Heine, B., and Narrog, H., eds., *Oxford Handbook of Linguistic Analysis*. OUP. 313–341.

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

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

Munro, R.; Bethard, S.; Kuperman, V.; Lai, V. T.; Melnick, R.; Potts, C.; Schnoebelen, T.; and Tily, H. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proc. Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*, 122–130.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 EMNLP*, 254–263. Honolulu, HI: ACL.