Crowdsourcing Annotation of Non-Local Semantic Roles

Parvin Sadat Feizabadi
Institut für Computerlinguistik
Heidelberg University
69120 Heidelberg, Germany
feizabadi@cl.uni-heidelberg.de

Sebastian Padó
Institut für Maschinelle Sprachverarbeitung
Stuttgart University
70569 Stuttgart, Germany
pado@ims.uni-stuttgart.de

Abstract

This paper reports on a study of crowdsourcing the annotation of *non-local* (or *implicit*) frame-semantic roles, i.e., roles that are realized in the previous discourse context. We describe two annotation setups (marking and gap filling) and find that gap filling works considerably better, attaining an acceptable quality relatively cheaply. The produced data is available for research purposes.

1 Introduction

In the last years, crowdsourcing, e.g., using Amazon's Mechanical Turk platform, has been used to collect data for a range of NLP tasks, e.g., MT evaluation (Callison-Burch, 2009), sentiment analysis (Mellebeek et al., 2010), and student answer rating (Heilman and Smith, 2010). Frame-semantic role annotation (FSRA) is a task that requires more linguistic expertise than most data collection tasks realized with crowdsourcing; nevertheless it is also a crucial prerequisite for high-performance frame-semantic role labeling (SRL) systems (Das et al., 2014). Thus, there are some studies that have investigated FSRA as a crowdsourcing task. It can be separated into two parts: First, choosing the frame evoked by a given predicate in a sentence; second, assigning the semantic roles associated with the chosen frame. Hong and Baker (2011) have recently addressed the first step, experimenting with various ways of presenting the task. Fossati et al. (2013) have considered both steps and operationalized them separately and jointly, finding the best results when a single annotation task is presented to turkers (due to the interdependence of the two steps) and when the semantic role description are simplified. Both studies conclude that crowdsourcing can produce usable results for FSRA but requires careful design. Our study extends these previous studies to the phenomenon of implicit (non-locally realized) semantic roles where annotators are presented with a target sentence in paragraph context, and have to decide for every role whether it is realized in the target sentence, elsewhere in the paragraph, or not at all. Our results shows that implicit roles can be annotated as well as locally realized roles in a crowdsourcing setup, again provided that good design choices are taken.

2 Implicit Semantic Roles

Implicit or non-locally realized semantic roles occur when arguments of a predicate are understood although not expressed in its direct syntactic neighborhood. FrameNet (Fillmore et al., 2003) distinguishes between indefinite non-instantiations (INIs), which are interpreted generically; definite non-instantiations (DNIs), which can often be identified with expressions from the previous context; and constructional non-instantiations (CNI), e.g., passives. For instance, in the following example, the GOAL of the predicate "reached" is realized locally, the SOURCE is a non-locally realized DNI, and the PATH is an INI and not realized at all.

(1) Phileas Fogg, having shut the door of [SOURCE his house] at half-past eleven, and having put his right foot before his left five hundred and seventy-five times, and his left foot before his right five hundred and seventy-six times, **reached** [GOAL the Reform Club].

Implicit roles play an important role in discourse comprehension and coherence (Burchardt et al., 2005) and have found increasing attention over the last years. The development was kickstarted by the creation of a corpus of non-local frame-semantic roles for the SemEval 2010 Task 10 (Ruppenhofer et al., 2010), which still serves as a de facto standard. A number of systems perform SRL for non-local roles (Chen et al., 2010; Silberer and Frank, 2012; Laparra and Rigau, 2013), but the obtained results are still far from satisfactory, with the best reported F-Score at 0.19. The main reason is data sparsity: Due to the small size of the dataset (just 438 sentences), every predicate occurs only a small number of times. Crowdsourcing can be an attractive strategy to acquire more annotations.

3 Experimental Setup

3.1 Domain

Our emphasis is on evaluating the annotation of implicit roles. We reduce complexity by limiting the number of frames and roles like earlier studies (Hong and Baker, 2011; Fossati et al., 2013). We focus on verbs from the MOTION and POSITION frames, which realize a common set of location roles (PLACE OF EVENT, SOURCE, GOAL, PATH). This makes the task more uniform and allows us to skip frame annotation. Information about spatial relations, provided by such verbs, can be useful for many NLP tasks which reason about spatial information, e.g. systems generating textual descriptions from visual data, robot navigation tasks, and geographical information systems or GIS (Kordjamshidi et al., 2012).

3.2 Corpus

We chose the novel "Around the World in Eighty Days" by Jules Verne, annotating the ten most frequent predicates meeting the conditions described above for annotation (*reach, arrive, descend, rush, follow, approach, send, cross, escape, pass*). A post-hoc analysis later showed that each instance of these predicates has on average 0.67 implicit roles identifiable in previous context, which underlines the relevance of annotating such cases. Metaphorical uses were discarded before annotation, which left an average 38.4 instances for each predicate.

4 Annotation and Agreement

We decided to present target sentences with three sentences of previous context, as a compromise between reading overhead and coverage of non-local roles: For nominalizations, the three previous sentences cover over 85% of all non-local roles (Gerber and Chai, 2012). An example and the detailed description of the task were provided to the annotators through external links. We experimented with two alternatives: annotation as a *marking task* and as a *gap filling task* (explained below). Each HIT was annotated by five turkers who were asked to annotate both local and non-local roles, since identification of local roles is necessary for reliable tagging of non-local roles.

	Source	Goal	Path	Place
Exact Match	0.35	0.44	0.48	0.24
Overlap	0.35	0.46	0.52	0.27

Table 1: Raw agreement among annotators in the "marking" task

4.1 Marking Task

Our rationale was to make the task as comprehensible as possible for non-experts. In each HIT, the target predicate in its context was shown in boldface and the annotators were asked to answer four questions about "the event in bold": (a) where does the event take place?; (b) what is its starting point?; (c) what is its end point?; (d) which path is used? For every question, turkers were asked to either mark a text span (shown in a non-editable field below the question) or click a button labeled "not found in the text". The goals of this setup were (a) to minimize annotation effort, and (b) to make the task as layman-compatible as possible, following Fossati et al.'s (2013) observation that linguistic definitions can harm results.

After annotating some instances, we computed raw inter-annotator agreement (IAA). Table 1 shows IAA among turkers in two conditions (average pairwise Exact Match and word-based Overlap) overall annotations for the first 49 instances.[1] The overall IAA is 37.9% (Exact Match) and 40.1% (Overlap). We found these results to be too low to continue this approach. The low results for Overlap indicate that the problems cannot be due mainly to differences in the marked spans. Indeed, an analysis showed that the main reason was that annotators were often confused by the presence of multiple predicates in the paragraph. Consequently, many answers marked roles pertaining not to the bolded target predicate but to other predicates, such as (2).

(2) Leaving Bombay, it passes through Sal-

[1] Kappa is not applicable since we have a large number of disjoint annotators.

	Source	Goal	Path	Place
Exact Match	0.46	0.46	0.56	0.30
Overlap	0.50	0.54	0.58	0.38

Table 2: Raw agreement among annotators in the "gap filling" task

> cette, **crossing** to the continent opposite Tannah, goes over the chain of the Western Ghauts, [. . .] and, descending southeastward by Burdivan and the French town of Chandernagor, has its terminus at Calcutta.

Annotators would be expected to annotate *the continent opposite Tannah* as the goal of crossing, but some annotated *Calcutta*, the final destination of the chain of motion events described.

4.2 Gap Filling Task

Seeing that the marking task did not constrain the interpretation of the turkers sufficiently, we moved to a second setup, gap filling, with the aim of focussing the turkers' attention to a single predicate rather than the complete set of predicates present in the text shown. In this task, the annotators were asked to complete the sentence by filling in the blanks in two sentences:

1. [Agent] [Event+ed] from . . . to . . . through . . . path.
2. The whole event took place in/at . . .

The first sentence corresponds to annotations of the SOURCE, GOAL, and PATH roles; the second one of the PLACE role. The rationale is that the presence of the predicate in the sentence focuses the turkers' attention on the predicate's actual roles. Annotators could leave gaps empty (in the case of unrealized roles), and we asked them to remain as close to the original material as possible, that is, avoid paraphrases. Perfect copying is not always possible, due to grammatical constraints.

Table 2 shows the IAA for this design. We see that even though the gap filling introduced a new source of variability (namely, the need for annotators to copy text), the IAA improves considerably, by up to 11% in Exact Match and 15% in Overlap. The new overall IAAs are 44.7% (+6.8%) and 50.2% (+10.1%), respectively. Overall, the numbers are still fairly low. However, note that these IAA numbers among turkers are a lower bound for the agreement between a "canonical" version of the turkers' annotation (see Section 5) and an ideal gold standard. Additionally, a data analysis showed that in the gap filling setup, many of the disagreements are more well-behaved: unsurprisingly, they are often cases where annotators disagree on the exact range of the string to fill into the gap. Consider the following example:

(3) Skillful detectives have been **sent** to all the principal ports of America and the Continent, and he'll be a clever fellow if he slips through their fingers."

Arguably, experts would annotate *all the principal ports of America and the Continent* as the GOAL role of **sent**. Turkers however annotated different spans, including *all the principal ports of America*, *ports*, as well as the "correct" span. The lowest IAA is found for the place role. While it is possible that our setup which required turkers to consider a second sentence to annotate place contributes to the overall difficulty, our data analysis indicates that the main problem is the more vague nature of PLACE compared to the other roles which made it more difficult for annotators to tag consistently. Consider Example (1): the PLACE could be, among other things, *the City*, *London*, *England*, etc. The large number of locations in the novel is a compounding factor. We found that for some predicates (e.g. *arrive, reach*), many turkers attempted to resolve the ambiguity by (erroneously) annotating the same text as both GOAL and PLACE, which runs counter to the FrameNet guidelines.

5 Canonicalization

We still need to compute a "canonical" annotation that combines the five turker's annotations. First, we need to decide whether a role should be realized or left unrealized (i.e., INI, CNI, or DNI but not in the presented context). Second, we need to decide on a span for realized roles. Canonicalization in crowdsourcing often assumes a majority principle, accepting the analysis proposed by most turkers. We found it necessary to be more flexible. Regarding realization, a manual analysis of a few instances showed that cases of two turker annotations with non-empty overlap could be accepted as non-local roles. That is, turkers frequently miss non-local roles, but if two out of five annotate an overlapping span with the same role, this is reasonable evidence. Regarding the role's span, we used the consensus

	Source	Goal	Path	Place
Exact Match	0.72	0.67	0.82	0.50
Overlap	0.72	0.69	0.82	0.54

Table 3: Raw agreement between canonical crowdsourcing annotation and expert annotation by role

	Local	Non-Local	Unrealized
Exact Match	0.66	0.66	0.69
Overlap	0.69	0.70	0.69

Table 4: Raw agreement between canonical annotation and expert annotation by realization status

span if it existed, and the maximal (union) span otherwise, given that some turkers filled the gaps just with head words and not complete constituents. To test the quality of the canonical annotation, one of the authors had previously annotated 100 random instances that were also presented to the turkers. We consider the result to be an expert annotation approximating a gold standard and use it to judge the quality of the canonical turker annotations. The results are shown in Table 3.

The overall raw agreement numbers are 67.80% (Exact Match) and 69.34% (Overlap). As we had hoped, the agreement between the canonical crowdsourcing annotation and the expert annotation is again substantially higher than the IAA among turkers. Again, we see the highest numbers for path (the most specific role) and the lowest numbers for place (the least specific role).

To assess whether the number obtained in table 3 are sensitive to realization status (explicit, implicit or unrealized), we broke down the agreement numbers by realization status. Somewhat to our (positive) surprise, the results in Table 4 indicate that non-locally realized roles are annotated ablut as reliably as locally realized ones. Except for the ill-defined PLACE role, our reliability is comparable to Fossati et al. (2013). Given the more difficult nature of the task (annotators are given more context and have to make a more difficult decision), we consider this a promising result.

6 Final Dataset and Cost

The final dataset consists of 384 predicate instances.[2] With four roles per predicate, a total of 1536 roles could have been realized. We found that more than half (60%) of the roles remained unrealized even in context. 23% of the roles were realized locally, and 17% non-locally. The distribution over locally realized, non-locally realized, and unrealized roles varies considerably among the four roles that we consider. GOAL has the highest percentage of realized roles overall (unrealized only for 34% of all predicate instances), and at the same time the highest ratio of locally realized roles (48% locally realized, 18% non-locally). This corresponds well to FrameNet's predictions about our chosen predicates which realize the Goal role generally as the direct object (*reach*) or an obligatory prepositional phrase (*arrive*). In contrast, SOURCE is realized only for 36% of all instances, and then predominantly non-locally (24% non-local vs. 12% local). This shows once more that a substantial part of predicate-argument structure must be recovered from previous discourse context.

On average, each HIT page was annotated in 1 minute and 48 seconds, which means 27 seconds per each role and a total of 60 hours for the whole annotation. We paid 0.15 USD for each HIT. Since the number of roles in all HITs was fixed to four (source, goal, path and place), each role cost 0.04 USD, which corresponds to about USD 0.19 for every canonical role annotation. This is about twice the amount paid by Fossati et al. and reflects the increased effort inherent in a task that involves discourse context.

7 Conclusion

This paper presented a study on crowdsourcing the annotation of non-local semantic roles in discourse context, comparing a marking and a gap filling setup. We found that gap filling is the more reliable choice since the repetition of the predicate helps focusing the turkers' attention on the roles at hand rather than understanding of the global text. Thus, the semantic role-based crowdsourcing approach of Fossati et al. (2013) appears to be generalizable to the area of non-locally realized roles, provided that the task is defined suitably. Our results also support Fossati et al.'s observation that reliable annotations can be obtained without providing definitions of semantic roles. However, we also find large differences among semantic roles. Some (like PATH) can be annotated reliably and should be usable to train or improve SRL systems. Others (like PLACE) are defined so vaguely that it is unclear how usable their annotations are.

[2] It can be downloaded for research purposes from `http://www.cl.uni-heidelberg.de/~feizabadi/res.mhtml`

References

Aljoscha Burchardt, Anette Frank, and Manfred Pinkal. 2005. Building text meaning representations from contextually related frames – a case study. In *Proceedings of the International Workshop on Computational Semantics*, pages 66–77, Tilburg, Netherlands.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267, Uppsala, Sweden.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*. To appear.

Charles J Fillmore, Christopher R Johnson, and Miriam R L Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Marco Fossati, Claudio Giuliano, Sara Tonelli, and Fondazione Bruno Kessler. 2013. Outsourcing FrameNet to the Crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 742–747, Sofia, Bulgaria.

Matthew Gerber and Joyce Y Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798.

Michael Heilman and Noah A Smith. 2010. Rating computer-generated questions with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 35–40, Los Angeles, CA.

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, Oregon, USA.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 365–373, Montréal, Canada.

Egoitz Laparra and German Rigau. 2013. Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 155–166, Potsdam, Germany.

Bart Mellebeek, Francesc Benavent, Jens Grivolla, Joan Codina, Marta R Costa-Jussa, and Rafael Banchs. 2010. Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 114–121, Los Angeles, CA.

Josef Ruppenhofer, Caroline Sporleder, R. Morante, Collin Baker, and Martha Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50, Uppsala, Sweden.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 1–10, Montreal, Canada.