Outsourcing FrameNet to the Crowd

Marco Fossati, Claudio Giuliano, and Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
{fossati,giuliano,satonelli}@fbk.eu

Abstract

We present the first attempt to perform full FrameNet annotation with crowdsourcing techniques. We compare two approaches: the first one is the standard annotation methodology of lexical units and frame elements in two steps, while the second is a novel approach aimed at acquiring frames in a bottom-up fashion, starting from frame element annotation. We show that our methodology, relying on a single annotation step and on simplified role definitions, outperforms the standard one both in terms of accuracy and time.

1 Introduction

Annotating frame information is a complex task, usually modeled in two steps: first annotators are asked to choose the situation (or *frame*) evoked by a given predicate (the *lexical unit, LU*) in a sentence, and then they assign the semantic roles (or *frame elements, FEs*) that describe the participants typically involved in the chosen frame. Existing frame annotation tools, such as Salto (Burchardt et al., 2006) and the Berkeley system (Fillmore et al., 2002) foresee this two-step approach, in which annotators first select a frame from a large repository of possible frames (1,162 frames are currently listed in the online version of the resource), and then assign the FE labels constrained by the chosen frame to LU dependents.

In this paper, we argue that such workflow shows some redundancy which can be addressed by radically changing the annotation methodology and performing it in one single step. Our novel annotation approach is also more compliant with the definition of *frames* proposed in Fillmore (1976): in his seminal work, Fillmore postulated that the meanings of words can be understood on the basis of a semantic frame, i.e. a description of a type of event or entity and the participants in it. This implies that frames can be distinguished one from another on the basis of the participants involved, thus it seems more cognitively plausible to start from the FE annotation to identify the frame expressed in a sentence, and not the contrary.

The goal of our methodology is to provide full frame annotation in a single step and in a bottom-up fashion. Instead of choosing the frame first, we focus on FEs and let the frame emerge based on the chosen FEs. We believe this approach complies better with the cognitive activity performed by annotators, while the 2-step methodology is more artificial and introduces some redundancy because part of the annotators' choices are replicated in the two steps (i.e. in order to assign a frame, annotators implicitly identify the participants also in the first step, even if they are annotated later).

Another issue we investigate in this work is how semantic roles should be annotated in a crowdsourcing framework. This task is particularly complex, therefore it is usually performed by expert annotators under the supervision of linguistic experts and lexicographers, as in the case of FrameNet. In NLP, different annotation efforts for encoding semantic roles have been carried out, each applying its own methodology and annotation guidelines (see for instance Ruppenhofer et al. (2006) for FrameNet and Palmer et al. (2005) for PropBank). In this work, we present a pilot study in which we assess to what extent role descriptions meant for 'linguistics experts' are also suitable for annotators from the crowd. Moreover, we show how a simplified version of these descriptions, less bounded to a specific linguistic theory, improve the annotation quality.

2 Related work

The construction of annotation datasets for NLP tasks via non-expert contributors has been ap-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 742–747,
Sofia, Bulgaria, August 4-9 2013.

proached in different ways, the most prominent being games with a purpose (GWAP) and micro-tasks. Verbosity (Von Ahn et al., 2006) was one of the first attempts in gathering annotations with a GWAP. Phrase Detectives (Chamberlain et al., 2008; Chamberlain et al., 2009) was meant to gather a corpus with coreference resolution annotations. Snow et al. (2008) described design and evaluation guidelines for five natural language micro-tasks. However, they explicitly chose a set of tasks that could be easily understood by non-expert contributors, thus leaving the recruitment and training issues open. Negri et al. (2011) built a multilingual textual entailment dataset for statistical machine translation systems.

The semantic role labeling problem has been recently addressed via crowdsourcing by Hong and Baker (2011). Furthermore, Baker (2012) highlighted the crucial role of recruiting people from the crowd in order to bypass the need for linguistics expert annotations. Nevertheless, Hong and Baker (2011) focused on the frame discrimination task, namely selecting the correct frame evoked by a given lemma. Such task is comparable to the word sense disambiguation one as per (Snow et al., 2008), although the complexity increased, due to lower inter-annotator agreement values.

3 Experiments

In this section, we describe the anatomy and discuss the results of the tasks we outsourced to the crowd via the CrowdFlower[1] platform.

Golden data Quality control of the collected judgements is a key factor for the success of the experiments. Cheating risk is minimized by adding *gold* units, namely data for which the requester already knows the answer. If a worker misses too many gold answers within a given threshold, he or she will be flagged as untrusted and his or her judgments will be automatically discarded.

Worker switching effect Depending on their accuracy in providing answers to gold units, workers may switch from a trusted to an untrusted status and vice versa. In practice, a worker submits his or her responses via a web page. Each page contains one gold unit and a variable number of regular units that can be set by the requester during the calibration phase. If a worker becomes untrusted, the platform collects another judgment to fill the gap. If a worker moves back to the trusted status, his or her previous contribution is added to the results as free extra judgments. Such phenomenon typically occurs when the complexity of gold units is high enough to induce low agreement in workers' answers. Thus, the requester is constrained to review gold units and to eventually forgive workers who missed them. This has massively happened in our experiments and is one of the main causes of the overall cost decrease and time increase.

Cost calibration The total cost of a generic crowdsourcing task is naturally bound to a data unit. This represents an issue in most of our experiments, as the number of questions per unit (i.e. a sentence) varies according to the number of frames and FEs evoked by the LU contained in a sentence. In order to enable cost comparison, for each experiment we need to use the average number of questions per sentence as a multiplier to a constant cost per sentence. We set the payment per working page to 5 $ cents and the number of sentences per page to 3, resulting in 1.83 $ cent per sentence.

3.1 Assessing task reproducibility and worker behavior change

Since our overall goal is to compare the performance of FrameNet annotation using our novel workflow to the performance of the standard, 2-step approach, we first take into account past related works and try to reproduce them.

To our knowledge, the only attempt to annotate frame information through crowdsourcing is the one presented in Hong and Baker (2011), which however did not include FE annotation.

Modeling The task is designed as follows. (a) Workers are invited to read a sentence where a LU is bolded. (b) The question `Which is the correct sense?` is combined with the set of frames evoked by the given LU, as well as the `None` choice. Finally, (c) workers must select the correct frame. A set of example sentences corresponding to each possible frame is provided in the instructions to facilitate workers.

As a preliminary study, we wanted to assess to what extent the proposed task could be reproduced and if workers reacted in a comparable way over time. Hong and Baker (2011) did not publish the input datasets, thus we ignore which sen-

[1] `https://crowdflower.com`

LU	2013 Sentences (Gold)	2013 Accuracy	2011 Accuracy
high.a	68 (9)	91.8	92
history.n	72 (9)	84.6	86
range.n	65 (8)	95	93
rip.v	88 (12)	81.9	92
thirst.n	29 (4)	90.4	95
top.a	36 (5)	98.7	96

Table 1: Comparison of the reproduced frame discrimination task as per (Hong and Baker, 2011)

tences were used. Besides, the authors computed accuracy values directly from the results upon a majority vote ground truth. Therefore, we decided to consider the same LUs used in Hong and Baker's experiments, i.e. *high*.a, *history*.n, *range*.n, *rip*.v, *thirst*.n and *top*.a, but we leveraged the complete sets of FrameNet 1.5 expert-annotated sentences as gold-standard data for immediate accuracy computation.

Discussion Table 1 displays the results we achieved, jointly with the experiments by Hong and Baker (2011). For the latter, we only show accuracy values, as the number of sentences was set to a constant value of 18, 2 of which were gold. If we assume that the crowd-based ground truth in 2011 experiments is approximately equivalent to the expert one, workers seem to have reacted in a similar manner compared to Hong and Baker's values, except for *rip*.v.

3.2 General task setting

We randomly chose the following LUs among the set of all verbal LUs in FrameNet evoking 2 frames each: *disappear.v* [CEASING_TO_BE, DEPARTING], *guide.v* [COTHEME, INFLUENCE_OF_EVENT_ON_COGNIZER], *heap.v* [FILLING, PLACING], *throw.v* [BODY_MOVEMENT, CAUSE_MOTION]. We considered verbal LUs as they usually have more overt arguments in a sentence, so that we were sure to provide workers with enough candidate FEs to annotate. Linguistic tasks in crowdsourcing frameworks are usually decomposed to make them accessible to the crowd. Hence, we set the polysemy of LUs to 2 to ensure that all experiments are executed using the smallest-scale subtask. More frames can then be handled by just replicating the experiments.

3.3 2-step approach

After observing that we were able to achieve similar results on the frame discrimination task as in previous work, we focused on the comparison between the 2-step and the 1-step frame annotation approaches.

We first set up experiments that emulate the former approach both in frame discrimination and FEs annotation. This will serve as the baseline against our methodology. Given the pipeline nature of the approach, errors in the frame discrimination step will affect FE recognition, thus impacting on the final accuracy. The magnitude of such effect strictly depends on the number of FEs associated with the wrongly detected frame.

3.3.1 Frame discrimination

Frame discrimination is the first phase of the 2-step annotation procedure. Hence, we need to leverage its output as the input for the next step.

Modeling The task is modeled as per Section 3.1.

Discussion Table 2 gives an insight into the results, which confirm the overall good accuracy as per the experiments discussed in Section 3.1.

3.3.2 Frame elements recognition

We consider all sentences annotated in the previous subtask with the frame assigned by the workers, even if it is not correct.

Modeling The task is presented as follows. (a) Workers are invited to read a sentence where a LU is bolded and the frame that was identified in the first step is provided as a title. (b) A list of FE definitions is then shown together with the FEs text chunks. Finally, (c) workers must match each definition with the proper FE.

Simplification Since FEs annotation is a very challenging task, and FE definitions are usually meant for experts in linguistics, we experimented with three different types of FE definitions: the original ones from FrameNet, a manually simplified version, and an automatically simplified one, using the tool by Heilman and Smith (2010). The latter simplifies complex sentences at the syntactic level and generates a question for each of the extracted clauses. As an example, we report below three versions obtained for the *Agent* definition in the DAMAGING frame:

Approach	2-STEP		1-STEP
Task	FD	FER	
Accuracy	.900	.687	**.792**
Answers	100	160	416
Trusted	100	100	84
Untrusted	21	36	217
Time (h)	102	69	**130**
Cost/question ($ cents)	1.83	2.74	8.41

Table 2: Overview of the experimental results. FD stands for Frame Discrimination, FER for FEs Recognition

Original: The conscious entity, generally a person, that performs the intentional action that results in the damage to the Patient.

Manually simplified: This element describes the person that performs the intentional action resulting in the damage to another person or object.

Automatic system: What that performs the intentional action that results in the damage to the Patient?

Simplification was performed by a linguistic expert, and followed a set of straightforward guidelines, which can be summarized as follows:

- When the semantic type associated with the FE is a common concept (e.g. `Location`), replace the FE name with the semantic type.
- Make syntactically complex definitions as simple as possible.
- Avoid variability in FE definitions, try to make them homogeneous (e.g. they should all start with "This element describes..." or similar).
- Replace technical concepts such as `Artifact` or `Sentient` with common words such as `Object` and `Person` respectively.

Although these changes (especially the last item) may make FE definitions less precise from a lexicographic point of view (for instance, sentient entities are not necessarily persons), annotation became more intuitive and had a positive impact on the overall quality.

After few pilot annotations with the three types of FE definitions, we noticed that the simplified one achieved a better accuracy and a lower number of untrusted annotators compared to the others. Therefore, we use the simplified definitions in both the 2-step and the 1-step approach (Section 3.4).

Discussion Table 2 provides an overview of the results we gathered. The total number of answers differs from the total number of trusted judgments, since the average value of questions per sentence amounts to 1.5.[2] First of all, we notice an increase in the number of untrusted judgments. This is caused by a generally low inter-worker agreement on gold sentences due to FE definitions, which still present a certain degree of complexity, even after simplification. We inspected the full reports sentence by sentence and observed a propagation of incorrect judgments when a sentence involves an unclear FE definition. As FE definitions may mutually include mentions of other FEs from the same frame, we believe this circularity generated confusion.

3.4 1-step approach

Having set the LU polysemy to 2, in our case a sentence S always contains a LU with 2 possible frames (f_1, f_2), but only conveys one, e.g. f_1. We formulate the approach as follows. S is replicated in 2 data units (S_a, S_b). Then, S_a is associated to the set E_1 of f_1 FE definitions, namely the correct ones for that sentence. Instead, S_b is associated to the set E_2 of f_2 FE definitions. We call S_b a *cross-frame* unit. Furthermore, we allow workers to select the `None` answer. In practice, we ask a total amount of $|E_1 \cup E_2| + 2$ questions per sentence S. In this way, we let the frame directly emerge from the FEs. If workers correctly answer `None` to a FE definition $d \in E_2$, the probability that S evokes f_1 increases.

Modeling Figure 1 displays a screenshot of the worker interface. The task is designed as per Section 3.3.2, but with major differences with respect to its content. This is better described by an example. The sentence `Karen threw her arms round my neck, spilling champagne everywhere` contains the LU *throw*.v evoking the frame BODY_MOVEMENT. However, *throw*.v is ambiguous and may also evoke CAUSE_MOTION. We ask to annotate both the BODY_MOVEMENT and the CAUSE_MOTION

[2] Cf. Section 3 for more details

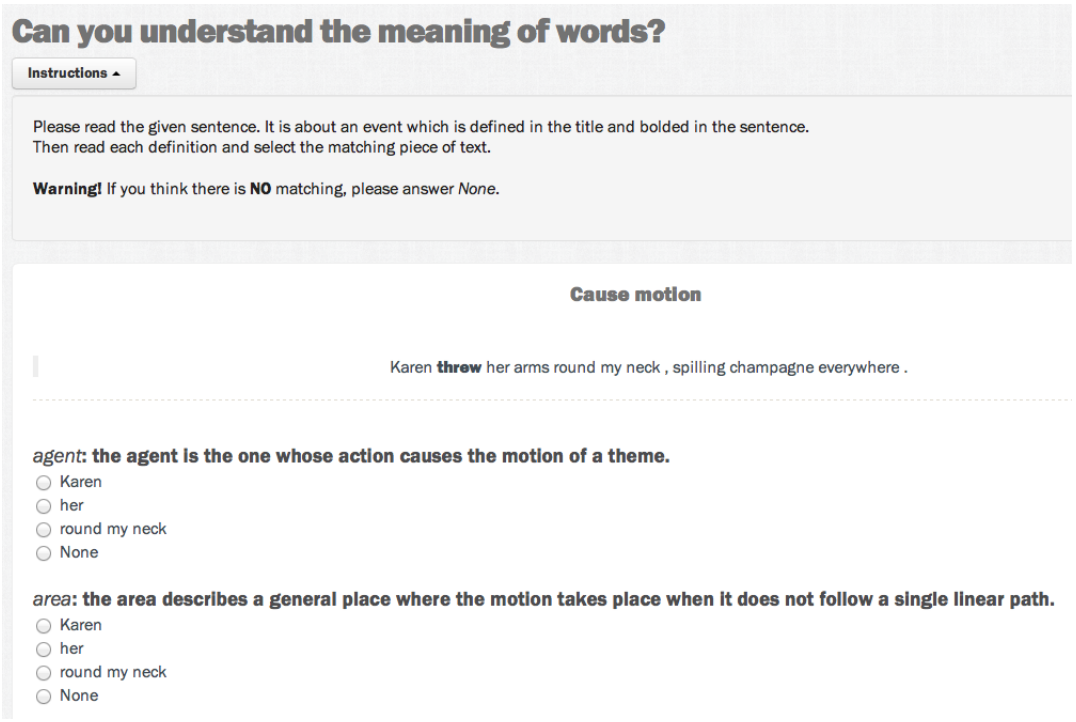

Figure 1: 1-step approach worker interface

core FEs, respectively as regular and cross-frame units.

Discussion We do not interpret the `None` choice as an abstention from judgment, since it is a correct answer for cross-frame units. Instead of precision and recall, we are thus able to directly compute workers' accuracy upon a majority vote. We envision an improvement with respect to the 2-step methodology, as we avoid the proven risk of error propagation originating from wrongly annotated frames in the first step. Table 2 illustrates the results we collected. As expected, accuracy reached a consistent enhancement. This demonstrates the hypothesis we stated in Section 1 on the cognitive plausibility of a bottom-up approach for frame annotation. Furthermore, the execution time decreases compared to the sum of the 2 steps, namely 130 hours against 171. Nevertheless, the cost is sensibly higher due to the higher number of questions that need to be addressed, in average 4.6 against 1.5. Untrusted judgments seriously grow, mainly because of the cross-frame gold complexity. Workers seem puzzled by the presence of `None`, which is a required answer for such units. If we consider the English FrameNet annotation agreement values between experts reported by Padó and Lapata (2009) as the upper bound (i.e., .897 for frame discrimination and .949 for FEs recognition), we believe our experimental setting can be reused as a valid alternative.

4 Conclusion

In this work, we presented an approach to perform frame annotation with crowdsourcing techniques, based on a single annotation step and on manually simplified FE definitions. Since the results seem promising, we are currently running larger scale experiments with the full set of FrameNet 1.5 annotated sentences. Input data, interface screenshots and full results are available and regularly updated at `http://db.tt/gu2Mj98i`.

Future work will include the investigation of a frame assignment strategy. In fact, we do not take into account the case of conflicting FE annotations in cross-frame units. Hence, we need a confidence score to determine which frame emerges if workers selected contradictory answers in a subset of cross-frame FE definitions.

Acknowledgements

The research leading to this paper was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

References

[Baker2012] Collin F Baker. 2012. Framenet, current collaborations and future goals. *Language Resources and Evaluation*, pages 1–18.

[Burchardt et al.2006] Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. Salto–a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, pages 517–520. Citeseer.

[Chamberlain et al.2008] Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. *Proceedings of I-Semantics, Graz*.

[Chamberlain et al.2009] Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2009. Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 57–62. Association for Computational Linguistics.

[Fillmore et al.2002] Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. The FrameNet Database and Software Tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1157–1160, Las Palmas, Spain.

[Fillmore1976] Charles J. Fillmore. 1976. Frame Semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language*, pages 20–32. Blackwell Publishing.

[Heilman and Smith2010] Michael Heilman and Noah A. Smith. 2010. Extracting Simplified Statements for Factual Question Generation. In *Proceedings of QG2010: The Third Workshop on Question Generation*, Pittsburgh, PA, USA.

[Hong and Baker2011] Jisup Hong and Collin F Baker. 2011. How good is the crowd at "real" wsd? *ACL HLT 2011*, page 30.

[Negri et al.2011] Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 670–679, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Padó and Lapata2009] Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

[Palmer et al.2005] Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, 31(1).

[Ruppenhofer et al.2006] Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. Available at http://framenet.icsi.berkeley.edu/book/book.html.

[Snow et al.2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

[Von Ahn et al.2006] Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM.