Crowdsourcing a Large Dataset of Domain-Specific Context-Sensitive Semantic Verb Relations

Maria Sukhareva[†], Judith Eckle-Kohler[†], Ivan Habernal[†], Iryna Gurevych[†‡]
† Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany
‡ Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information, Schloßstr. 29, 60486 Frankfurt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

We present a new large dataset of 12403 context-sensitive verb relations manually annotated via crowdsourcing. These relations capture fine-grained semantic information between verb-centric propositions, such as temporal or entailment relations. We propose a novel semantic verb relation scheme and design a multi-step annotation approach for scaling-up the annotations using crowdsourcing. We employ several quality measures and report on agreement scores. The resulting dataset is available under a permissive CreativeCommons license at `www.ukp.tu-darmstadt.de/data/verb-relations/`. It represents a valuable resource for various applications, such as automatic information consolidation or automatic summarization.

Keywords: crowdsourcing, semantic relations, new resource

1. Introduction

Automatic information consolidation (Saracevic, 1986) in the sense of supporting users in identifying and assimilating a large set of relevant statements found across multitudes of heterogeneous documents has recently gained attention (Wurzinger, 2010). Semantic relations between text segments play a key role in helping users to explore and navigate in the consolidated information. Often, the semantic relation between two segments is triggered by lexical-semantic relations between word senses in the two text segments. In our work, we consider the particular domain of everyday educational topics. This domain contains rather generic language, i.e., has low coverage of named entities. Consider the following example where the lexical-semantic *happens-before* relation between the verb senses *experience an emotion* and *report an emotion* triggers a *happens-before* relation between the two sentences (1) and (2).

> (1) Students with learning and behavioral disabilities are more likely to experience school stress.
>
> (2) Students with learning and behavioral disabilities report school stress as the primary reason for drug use.

A system for information consolidation could strongly benefit from a resource of lexical semantic relations between verb senses, such as the *happens-before* relation given above. This kind of information would help the system to reveal an inference between the two sentences (1) and (2) which would otherwise remain hidden, because it is not expressed on the text surface via linguistic means, such as verb tense or discourse connectives. Apart from the temporal *happens-before* relation between verb senses, there are other relations which are important for information consolidation, e.g., the *cause* relation.

Although WordNet (Fellbaum, 1998) contains semantic relations between verb senses, it has a very limited coverage of semantic verb relations, in particular for domain-specific text (Levy et al., 2014). While various approaches for the automatic acquisition of lexical-semantic relations between verbs from large-scale corpora have been developed (e.g., (Lin and Pantel, 2001; Chklovski and Pantel, 2004; Hashimoto et al., 2009)), the extracted relations are not specified on the sense level, but rather on the word level. For instance, Hashimoto et al. (2009) used templates with variables to represent verb relations, which, however, lack semantic information about the verb context, and thus also about the verb sense.

We would like to emphasize that it is essential for resources of lexical-semantic verb relations to specify the relations on the word sense level, because verbs are highly polysemous. This can be achieved, for example, by using a context-sensitive representation that captures semantic information about the verb context. An example of such a context-sensitive formulation is shown in the next example: (3) and (4) use triples consisting of subject, verb and object, which we call *propositions*, to represent the lexical-semantic *happens-before* relation between the two verb senses in the examples (1) and (2) above:

> (3) (Students, experience, school stress)
>
> (4) (students, report, school stress)

In this work, we therefore aim at constructing a large and representative dataset of context-sensitive verb relations by means of crowdsourcing. We go beyond previous work on extracting verb relations from corpora and propose a novel semantic verb relation scheme designed for the challenging task of automatic information consolidation in the particular domain of everyday educational topics. In particular, we make the following contributions:

- We developed a linguistically motivated hierarchical annotation scheme of fine-grained semantic verb relations between propositions (Section 3.), including temporal, as well as causal and fine-grained entailment relations;
- we developed a sampling approach for extracting

proposition pairs that are candidates for the annotation of semantic relations;

- we designed an annotation workflow that is suitable for crowdsourcing by decomposing the annotation of semantic relations into hierarchically organized atomic judgments.

The resulting dataset, which we will make available under a permissive CC license, will not only be useful for automatic information consolidation, but also in related applications such as automatic summarization.

2. Related Work

Regarding lexical-semantic relations between verbs, there is prior work in the context of recognizing textual entailment (RTE) between two text segments (Dagan et al., 2006). Large resources of lexical entailments (so-called lexical entailment rules) are commonly used in RTE systems (Magnini et al., 2014), because the textual entailment between two segments is often triggered by lexical-semantic relations between word senses in the two segments. However, previous approaches for the automatic acquisition of lexical verb entailments have ignored the semantic context of the verbs (Lin and Pantel, 2001; Chklovski and Pantel, 2004; Hashimoto et al., 2009)), and thus also their word sense; word sense information is crucial especially for verbs due to their high polysemy.

In recent years, crowdsourcing proved to be a fast and efficient way to address various linguistic annotation tasks that have traditionally been solved by trained linguists. The OpenCorpora project has gathered a large amount of morphological and morphosyntactic annotations for Russian (Ustalov, 2014). Chamberlain et al. (2008) produced a corpus annotated with co-reference resolutions. (Feizabadi and Padó, 2014) used crowdsourcing to create a dataset for semantic role labeling. Fossati et al. (2013) describes a crowdsourcing technique to produce FrameNet annotation. Snow et al. (2008) investigated the applicability of crowdsourcing to five semantic NLP tasks: word similarity, event annotation, word sense disambiguation, sentiment analysis and textual entailment; they concluded that crowdsourcing was well-suitable for these tasks.

In particular, crowdsourcing of paraphrase and textual entailment annotations came into focus of multiple recent studies. Negri et al. (2011) built a cross-lingual textual entailment corpus by outsourcing the annotation tasks to *Amazon Mechanical Turk* and introduced the methodology of annotation task decomposition into simple subtasks. Similarly, Zeller and Padó (2013) created a large dataset for German textual entailment by dividing the annotation task into three separate phases: summarization, paraphrasing and validation. Splitting a complex annotation task into a sequence of easily explainable microtasks was also successfully applied by Zeichner et al. (2012) for evaluating lexical inference rules.

In our research we also decompose the task of annotating lexical semantic relations between propositions into a series of *atomic judgments*. Unlike aforementioned approaches, we do not restrain ourselves to one specific semantic relation, such as the quite general (textual) entailment relation, but introduce a crowdsourcing methodology for annotating a range of more fined-grained semantic relations between propositions.

3. Annotation Scheme

We developed an annotation scheme for semantic relations between propositions. Propositions describe events, states or actions; we will use the term event in the rest of this section to refer to all of them. Propositions can describe the same event, which we call a co-reference of propositions. Figure 1 gives an overview of this new hierarchical annotation scheme.

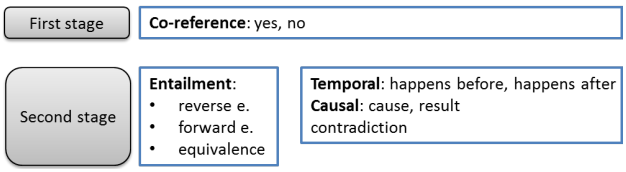

Figure 1: Annotation scheme for semantic relations between propositions.

Based on the lexical-semantic relations between verbs in WordNet (Fellbaum, 1998), we consider three major groups of semantic verb relations between propositions: *temporal*, *causal*, and *entailment*.

For the entailment relation, we follow Pavlick et al. (2015) and focus on the three types *forward entailment*, *reverse entailment* and *equivalence*. While these entailment relationships require that the related propositions co-refer (i.e. describe the same event; see figure 2, example 1), the temporal relations *precedence* and *succession* imply that the two propositions do not refer to the same event.

Thus, the first level of annotation, given two propositions, should be the decision, if the propositions refer to the same event. Depending on the outcome of this decision, the annotation either proceeds with the entailment relations or with the temporal and causal relations.

Temporal relations are closely connected to causal relations, since two events must be temporally aligned in order for a causal relation to hold (Tooley, 1997). Causal relations, including the *result* relation, can be triggered by precedence or succession (see example 2 in figure 2), but can also hold between two events that happen simultaneously (see example 3 in figure 2).

As for the contradiction relation, it is assumed that it is not possible to determine any temporal alignment of contradicting propositions.

4. Annotation Study

4.1. Sampling Approach

We used a limited amount of 1617 documents from 9 educational topics (e.g., parenting, preschool, higher education) for sampling candidate proposition pairs. Our sampling approach consists of two steps which will describe in the next two subsections: first, the extraction of propositions, and second, the generation of candidate pairs from the extracted propositions.

(1) a. one professor, forward, the email.
b. one professor, circulate, the email.

(2) a. some students, experience, school stress
b. some students, report, school stress

(3) a. all students, receive, adequate training and encouragement
b. all students, enjoy, adequate training and encouragement

Figure 2: Examples of different semantic relations from Section 3.

4.1.1. Extraction of Propositions

For the proposition extraction, we first preprocessed the document collection using components from DKPro Core (Eckart de Castilho and Gurevych, 2014), in particular, the Stanford Core NLP POS-tagger, lemmatizer, Named Entity Recognizer, and dependency parser, as well as a semantic field annotator for noun, verb and adjective tokens which is based on the UBY resource (Gurevych et al., 2012).
We build upon the results of dependency parsing and extract propositions from sentence fragments with a lexical verb (we do not consider auxiliary and modal verbs), where the parser has annotated at least one of the dependency types `nsubj` (or alternatively `nsubjpass`, `agent`), `dopbj`, and `prep_`$\star$. We use a proposition template with four fixed slots for the verb, for the subject argument, the object argument and the prepositional argument. This way, we extract propositions that consist of one to three nominal arguments; the absence of a subject, object or prepositional object is indicated in the extracted template as well. For each nominal argument, we keep the following information types: (1) its corresponding phrase, (2) its head word, (3) its dependency type and (4) the semantic field of its head word.
The following example shows the information available for the extracted propositions (the numbers refer to the information types above):

S: The candidates (1) [candidates (2),nsubj (3),person (4)]
V: complete
O: their diploma (1) [diploma (2),dobj (3),communication (4)]
PO: <empty>

4.1.2. Generation of Proposition Pairs

Our generation of pairs from the extracted propositions relies on the following central assumption:

> When two propositions are semantically related, there needs to be some *semantic overlap*.

For instance, the semantic overlap can arise from arguments in the two propositions that co-refer or that have the same head word. Since we are interested in semantic relations between verbs, we want to find propositions where the verbs are different, but there is a semantic overlap in any of its arguments (e.g., the subject, the direct object or the prepositional object).

Sampling proposition pairs for human annotation is far from trivial, as there should not only be a semantic overlap between promising candidates, but the two propositions presented for annotation should also sound natural and provide more context than just the argument head words alone. Consider the following example showing a proposition pair where the subject head words and the object head words are identical. We list two variants, with and without the full argument phrase included as context, in order to illustrate the difference in interpretability:

(4) a. older school children, bully, children
b. older school children, punish, children

a. children, bully, children
b. children, punish, children

Sampling of proposition pairs for annotation can therefore be split into two subtasks: first, recovering propositions with semantic overlap, and second, generating naturally sounding pairs of propositions from them.
An example of the first subtask is given in the work by Levy et al. (2014) who annotated entailment relations between pairs of propositions. They recovered propositions with identical subject and object head words in a domain-specific subset[1] of the Google syntactic n-grams (Goldberg and Orwant, 2013), a huge dataset which is based on a corpus of 3.5 million English books.
We also recover proposition pairs with this kind of overlap in our set of extracted propositions, but in addition, we consider three further kinds of semantic overlap. We specify these four kinds of semantic overlap in the following pseudo-code 1.

Algorithm 1 Types of semantic overlap

Given $prop_1, prop_2$ ▷ two propositions
$prop_i \leftarrow (v_i, h(S_i), h(O_i), h(PO_i)$ ▷ verb + head word for subject, object and prepositional object
if $h(S_1) = h(S_2)$ **and** $h(O_1) = h(O_2)$
then $overlapType \leftarrow$ **SubjObj**
if $h(PO_1) = h(PO_2)$
then $overlapType \leftarrow$ **PrepObj**
if $h(O_1) = h(S_2)$ **and** $h(O_2).isEmpty$ ▷ second proposition has no object
then $overlapType \leftarrow$ **SubjIsObj21**
if $h(O_1) = h(S_2)$ **and not**$(h(O_2).isEmpty)$
then $overlapType \leftarrow$ **SubjIsObj22**

Regarding the second subtask, generating a naturally sounding pair of propositions for human annotation, we already demonstrated in our example 4 that adding context to the propositions by using the argument phrases instead of the head words is important. However, this can reduce the semantic overlap between two propositions considerably, thus making it harder for untrained annotators to recognize a semantic relation between the two verbs involved; this is illustrated in the next example (argument heads are underlined):

(5) a. parents, seek, information

[1] They considered the healthcare domain.

b. parents and professionals, need, more information

Therefore, we applied the following transformations to a pair of propositions with semantic overlap to generate more natural pairs for human annotation:

- For the overlap types *SubjObj* and *PrepObj*, we use the argument phrases from only *one* proposition to generate a pair: this way, the first item of the pair corresponds to the first original proposition, while the second item combines the arguments from the first proposition with the verb from the second original proposition. As a result, the second generated proposition is artificial, i.e. has not been seen in the data.
- For the overlap types *SubjIsObj21* and *SubjIsObj22*, we us the argument phrase of the object from the first original proposition as the argument phrase of the subject in the second (generated, i.e. not seen) proposition.

In example 6, we show a proposition pair with semantic overlap of type *SubjIsObj21*, first before the transformation and then after modifying the subject phrase of the second proposition. By using the object phrase from the first proposition also as subject phrase in the second proposition, we enforce a reading of the propositions where the overlapping subject and object co-refer.

(6) a. the parents, bring, the kids
b. kids, play (*possibly other kids*)
a. the parents, bring, the kids
b. the kids, play (*very likely the same kids*)

4.2. Crowdsourcing

The direct annotation of the sampled proposition pairs with lexical semantic relations was not feasible, because many propositions were not well-formed (e.g., containing incomplete noun phrases due to parsing errors, making them meaningless) Filtering out non-wellformed propositions is essential when propositions are automatically extracted, as reported also for instance by Bamman and Smith (2015). Hence, our overall annotation task was divided into two stages, first, the annotation of well-formedness for individual propositions and second, the annotation of semantic relations between propositions.

4.2.1. Annotating Well-formedness

In the first stage, the MTurkers[2] were asked to judge the well-formedness of a proposition given a short instruction. We assumed no linguistic knowledge of the MTurkers, so the guidelines contained no linguistic terminology. To ensure the MTurkers' reliability and their proficiency in the English language, their location was constrained to the US and their overall acceptance rate was required not to be lower than 95%.
In order to stimulate MTurkers to follow the guidelines, the guidelines were permanently displayed on the left side of the annotation page and were short enough to fit on the browser window without scrolling. Each HIT (Human Intelligence Task) contained fourteen propositions and required assignments from four MTurkers per proposition.
In total we obtained well-formedness judgments for 9,856 propositions (each with assignments from four MTurkers). The final well-formedness judgment was created using MACE[3], a state-of-the-art tool for automatic estimation of annotators' competence and of the most probable label. Finally, out of the 9,856 propositions, 4319 were labeled as well-formed (see Table 1).

Well-formed	Not well-formed	Annotators
4319(44%)	5537(56%)	4

Table 1: Well-formedness annotation; labels are estimated by MACE over labels assigned by four MTurkers.

4.2.2. Annotation of Semantic Relations

We used the well-formed propositions to filter the generated proposition pairs in each of the four semantic overlap types, see section 4.1.2.
Each proposition pair was assigned to four MTurkers. Rather than introducing a long qualification test that would examine the ability of a worker to consistently annotate lexical semantic relations, which would inevitably raise the costs of the annotations and make the task less attractive for the workers, we increased the requirements of the acceptance rate to 97% and provided a longer instruction.
MTurkers were asked to make a series of atomic judgments that constitute a decision tree (see Figure 3), as it turned out to be more feasible than judging the relation directly (which we tried during our pilot annotations).
In order to fit the guidelines on a browser page and ensure that MTurkers are not overwhelmed by its length, the guidelines were displayed gradually. Only the parts of the instruction that were relevant to the current atomic judgment were displayed to a MTurker. For example, if a MTurker was to make a decision on the temporal alignment, only the explanations on how to judge on precedence and succession were visible. MTurkers were also instructed to use various connectives to justify their decision: for example, the second proposition is more specific if it is possible to connect the propositions by *'specifically'*, *'in particular'*, or *'to be more precise'*[4].

Annotation Procedure On the top level, a MTurker was to decide whether the verbs refer to the same event, i.e., whether they co-refer. Depending on the MTurker's decision, she was prompted to either continue with the temporal alignment (no co-reference) or to move on to the annotation of entailment relations (if the two verbs co-referred), see Figure 3.

5. Results

In the pilot stage of the annotation project, we performed an evaluation of the annotation on a small batch of 176 proposition pairs. Table 3 summarizes the results in terms of percentage agreement and Krippendorff's α during the

[2] Amazon Mechanical Turk workers

[3] http://www.isi.edu/publications/licensed-sw/mace/

[4] An example of a HIT: `http://tinyurl.com/hnmp35e`

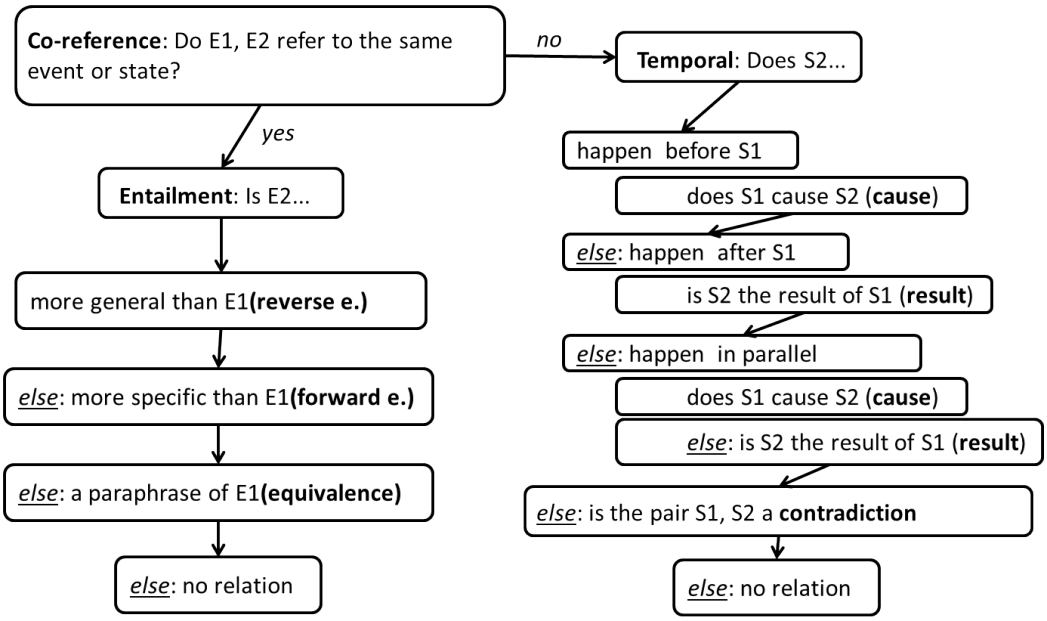

Figure 3: Hierarchy of atomic judgments for relation annotation; S_1, S_2 are situations (events) described by the textual expressions E_1, E_2 (used for brevity only in this figure, not in the annotation interface); relations in bold.

	SubjObj	SubjIsObj21	SubjIsObj22	PrepObj
Total	1664	3868	6437	434
Entailment	38%	26%	18%	44%
Temporal	62%	74%	82%	56%
Entailment relations				
reverse	7%	16%	4%	11%
forward	6%	3%	3%	6%
equivalence	13%	2%	2%	13%
none	9%	4%	7%	14%
Temporal relations				
precedence	15%	20%	25%	9%
succession	13%	7%	12%	7%
parallel	7%	7%	10%	8%
none	27%	40%	34%	24%
cause	2%	2%	2%	6%
contradiction	8%	5%	7%	6%
result	2%	9%	21%	2%
none	13%	5%	7%	7%

Table 2: Distribution of semantic relations in the final dataset. Numbers are given for each of the four semantic overlap types separately.

pilot annotations. The annotation of semantic relations proved to be a more challenging tasks for MTurkers than the well-formedness annotations. All in all, the MTurkers achieved a moderate level of agreement. For comparison, three experts also annotated well-formedness of 80 propositions and achieved 75% of agreement which corresponds to $\alpha = 0.51$.[5]

Krippendorff's α for the annotation of semantic relations drops with every level of annotation. The explanation lies in the unbalanced nature of the annotated data. Obviously, there are significantly fewer proposition pairs for which a causal relation holds rather than for which this relation does not hold. Also, it is important to keep in mind that the annotation categories are not mutually exclusive due to the hierarchical nature of the lexical semantic relations. For example, if the first event was annotated to cause the second then it might also hold that the first event precedes the second. Therefore, a chance-corrected agreement measure such as Krippendorff's α is less suitable for such data than the raw percentage agreement. The MTurkers showed a constant level of percentage agreement of over 70% for all the annotation categories with the best agreement on event co-reference and the worst on whether the second proposi-

[5]Although using Krippendorff's α as a measure for assessing quality of crowdsourcing is questionable (Ipeirotis et al., 2010; Aroyo and Welty, 2015), we use it here as a proxy to enable comparison with experts.

well-formedness		lexical semantic relations									
		Co-reference		Temporal		Cause		Result		Entailment	
%	α	%	α	%	α	%	α	%	α	%	α
72.33%	0.45	81.33%	0.45	76.38%	0.21	71.8%	0.32	70.15%	0.22	70.85%	0.27

Table 3: Inter-annotator agreement for individual semantic relations in the pilot study; observed agreement (%) and Krippendorff's α.

tion is the result of the first.

During the large-scale experiment we annotated 12403 proposition pairs in total. 44% of the proposition pairs were annotated by MTurkers as unrelated. The full overview of the dataset is given in Table 2. Similar as in the pilot annotations, the large dataset is imbalanced. Across all the semantic overlap types, temporal relations are prevailing over the entailment relations. The further relation distribution varies depending on the semantic overlap type: For the *SubjIsObj22* type, there are more causal relations which is consistent with the larger number of temporal relations overall.

The total cost of the whole crowdsourcing experiment, including well-formedness annotation and relation annotation was $4.410 USD.

6. Conclusion

This paper presented a semantic relations dataset for the domain of everyday educational topics.

On the initial stage of the annotation project, we conducted a pilot study which indicated that our multi-step annotation procedure is valid and that constructing the dataset by means of crowdsourcing is feasible. Inter-annotator agreement was comparable to related semantic annotation tasks that employ crowdsourcing.

Consequently, we annotated a large set of 9856 propositions regarding well-formedness and used the wellformed propositions to generate 12403 proposition pairs for the annotation of context-sensitive semantic verb relations. We developed a hierarchical annotation scheme for verb relations and applied it to the proposition pairs. The total cost of the annotation was 4,410 USD.

In conclusion, crowdsourcing a dataset of complex domain-specific context-sensitive verb relations can be a fast and inexpensive alternative to expert annotations. The dataset is publicly available under a permissive CC license.

We believe that the dataset is a valuable contribution to future work in domain-specific information consolidation or summarization.

7. Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant Nº I/82806, by the German Institute for Educational Research (DIPF) and by the German Research Foundation under grant No. GU 798/17-1. We would like to thank Omer Levy for discussions regarding the sampling approach.

8. Bibliographical References

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Bamman, D. and Smith, N. A. (2015). Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, Lisbon, Portugal, September. Association for Computational Linguistics.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Addressing the resource bottleneck to create large-scale annotated texts. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 375–380, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chklovski, T. and Pantel, P. (2004). VerbOcean: mining the web for fine-grained semantic verb relations. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, July. Association for Computational Linguistics.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Quinonero-Candela, J., Dagan, I., Magnini, B., and dAlche Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer Berlin Heidelberg.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Feizabadi, P. S. and Padó, S. (2014). Crowdsourcing annotation of non-local semantic roles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 226–230, Gothenburg, Sweden, April. Association for Computational Linguistics.

Fellbaum, C., editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Fossati, M., Giuliano, C., and Tonelli, S. (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.

Goldberg, Y. and Orwant, J. (2013). A dataset of syntactic-

ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247, Atlanta, Georgia, USA.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France, April. Association for Computational Linguistics.

Hashimoto, C., Torisawa, K., Kuroda, K., De Saeger, S., Murata, M., and Kazama, J. (2009). Large-scale verb entailment acquisition from the Web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1172–1181, Singapore, August. Association for Computational Linguistics.

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, page 64, New York, New York, USA. ACM Press.

Levy, O., Dagan, I., and Goldberger, J. (2014). Focused Entailment Graphs for Open IE Propositions. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 87–97, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Lin, D. and Pantel, P. (2001). DIRT: Discovery of Inference Rules from Text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 323–328, New York, NY, USA. ACM Press.

Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Padó, S., Stern, A., and Levy, O. (2014). The excitement open platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 43–48, Baltimore, Maryland, June. Association for Computational Linguistics.

Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., and Marchetti, A. (2011). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Pavlick, E., Bos, J., Nissim, M., Beller, C., Van Durme, B., and Callison-Burch, C. (2015). Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1512–1522, Beijing, China, July. Association for Computational Linguistics.

Saracevic, T. (1986). Processes and problems in information consolidation. *Information Processing & Management*, 22(1):45–60, January.

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 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October. Association for Computational Linguistics.

Tooley, M. (1997). *Time, Tense, and Causation*. Clarendon Press.

Ustalov, D. (2014). Towards crowdsourcing and cooperation in linguistic resources. *CoRR*, abs/1408.4245.

Wurzinger, G. (2010). Information consolidation in large bodies of information. *Journal of Universal Computer Science*, 16(21):3314–3323, dec.

Zeichner, N., Berant, J., and Dagan, I. (2012). Crowdsourcing inference-rule evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–160, Jeju Island, Korea, July. Association for Computational Linguistics.

Zeller, B. D. and Padó, S. (2013). A search task dataset for german textual entailment. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 288–299, Potsdam, Germany, March. Association for Computational Linguistics.