Adding Context to Semantic Data-Driven Paraphrasing

Vered Shwartz **Ido Dagan**
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
vered1986@gmail.com dagan@cs.biu.ac.il

Abstract

Recognizing lexical inferences between pairs of terms is a common task in NLP applications, which should typically be performed within a given context. Such context-sensitive inferences have to consider both term meaning in context as well as the fine-grained relation holding between the terms. Hence, to develop suitable lexical inference methods, we need datasets that are annotated with fine-grained semantic relations in-context. Since existing datasets either provide out-of-context annotations or refer to coarse-grained relations, we propose a methodology for adding context-sensitive annotations. We demonstrate our methodology by applying it to phrase pairs from PPDB 2.0, creating a novel dataset of *fine-grained* lexical inferences *in-context* and showing its utility in developing context-sensitive methods.

1 Introduction

Recognizing lexical inference is an essential component in semantic tasks. In question answering, for instance, identifying that *broadcast* and *air* are synonymous enables answering the question "When was 'Friends' first aired?" given the text "'Friends' was first broadcast in 1994". Semantic relations such as synonymy *(tall, high)* and hypernymy *(cat, pet)* are used to infer the meaning of one term from another, in order to overcome lexical variability.

In semantic tasks, such terms appear within corresponding contexts, thus making two aspects necessary in order to correctly apply inferences: First, the meaning of each term should be considered within its context (Szpektor et al., 2007; Pantel et al., 2007), e.g., *play* entails *compete* in certain contexts, but not in the context of playing the national anthem at a sports competition. Second, the soundness of inferences within context is conditioned on the fine-grained semantic relation that holds between the terms, as studied within natural logic (MacCartney and Manning, 2007). For instance, in upward-monotone sentences a term entails its hypernym ("my *iPhone*'s battery is low" ⇒ "my *phone*'s battery is low"), while in downward monotone ones it entails its hyponym ("talking on the *phone* is prohibited" ⇒ "talking on the *iPhone* is prohibited").

Accordingly, developing algorithms that properly apply lexical inferences in context requires datasets in which inferences are annotated *in-context* by *fine-grained* semantic relations. Yet, such a dataset is not available (see 2.1). Most existing datasets provide *out-of-context* annotations, while the few available *in-context* annotations refer to coarse-grained relations, such as relatedness or similarity.

In recent years, the PPDB paraphrase database (Ganitkevitch et al., 2013) became a popular resource among semantic tasks, such as monolingual alignment (Sultan et al., 2014) and recognizing textual entailment (Noh et al., 2015). Recently, Pavlick et al. (2015) classified each paraphrase pair to the fine-grained semantic relation that holds between the phrases, following natural logic (MacCartney and Manning, 2007). To that end, a subset of PPDB paraphrase-pairs were manually annotated, forming a fine-grained lexical inference dataset. Yet, annotations are given *out-of-context*, limiting its utility.

In this paper, we aim to fill the current gap in the inventory of lexical inference datasets, and present a methodology for adding context to out-of-context datasets. We apply our methodology on a subset of phrase pairs from Pavlick et al. (2015),

*Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 108–113,
Berlin, Germany, August 11-12, 2016.

	x	y	contexts	out-of-context relation	in-context relation
1	piece	strip	Roughly 1,500 gold and silver **pieces** were found and the hoard contains roughly 5kgs of gold and 2.5kgs of silver. A huge political storm has erupted around Australia after labor leader Kevin Rudd was found to have gone to a **strip** club during a taxpayer funded trip.	Equivalence	Independent
2	competition	race	Three countries withdrew from the **competition**: Germany, Spain and Switzerland. Morgan Tsvangirai, the leader of the Movement for Democratic Change (MDC), Zimbabwe's main opposition party, has said that he will pull out of the **race** to become the president of Zimbabwe.	Reverse Entailment	Equivalence
3	boy	family	The birth of the **boy**, whose birth name is disputed among different sources, is considered very important in the entertainment world. Bill will likely disrupt the Obama **family**'s vacation to Martha's Vineyard.	Forward Entailment	Other-related
4	jump	walk	Amid wild scenes of joy on the pitch he **jumped** onto the podium and lifted the trophy, the fourth of Italy's history. In a game about rescuing hostages a hero might **walk** past Coca-Cola machine's one week and Pepsi the next.	Other-related	Alternation

Table 1: Illustration of annotation shifts when context is given. [1] the sense of *strip* in the given context is different from the one which is equivalent to *piece*. [2] the term *race* is judged out-of-context as more specific than *competition*, but is considered equivalent to it in a particular context. [3] a meronymy relation is (often) considered out-of-context as entailment, while in a given context this judgment doesn't hold. [4] general relations may become more concrete when the context is given.

creating a novel dataset for fine-grained lexical inference in-context. For each term-pair, we add a pair of context sentences, and re-annotate these term-pairs with respect to their contexts.[1] We show that almost half of the semantically-related term-pairs become unrelated when the context is specified. Furthermore, a generic out-of-context relation may change within a given context (see table 1). We further report baseline results that demonstrate the utility of our dataset in developing fine-grained context-sensitive lexical inference methods.

2 Background

2.1 Lexical Inference Datasets

Figure 1 lists prominent human-annotated datasets used for developing lexical inference methods. In these datasets, each entry consists of an (x, y) term-pair, annotated to whether a certain semantic relation holds between x and y. Each dataset either specifies fine-grained semantic relations (see 2.2), or groups several semantic relations under a single coarse-grained relation (e.g. lexical substitution, similarity).

In some datasets, term-pairs are annotated to whether the relation holds between them in some (unspecified) contexts (*out-of-context*), while in others, the annotation is given with respect to a given context (*in-context*). In these datasets, each entry consists of a term-pair, x and y, and context, where some of the datasets provide a single context in which x occurs while others provide a separate context for each of x and y (corresponding to the *1 context* and *2 contexts* columns in Figure 1). The latter simulates a frequent need in NLP applications, for example, a question answering system recognizes that *broadcast* entails *air* given the context of the question ("When was 'Friends' first aired?") and that of the candidate passage ("'Friends' was first broadcast in 1994").

We observe that most lexical inference datasets provide *out-of-context* annotations. The existing *in-context* datasets are annotated for coarse-grained semantic relations, such as similarity or relatedness, which may not be sufficiently informative.

[1] The dataset and annotation guidelines are available at: `http://u.cs.biu.ac.il/~nlp/resources/downloads/context-sensitive-fine-grained-dataset`.

relation	none	1 context	2 contexts
fine-grained	[15] PPDB-fine-human		[16] this paper
coarse-grained	[1] WordSim-353 similarity [2] SimLex-999 [4] Annotated-PPDB [5] WordSim-353 relatedness [6] MEN [10] Kotlerman2010 [11] Turney2014 [12] Levy2014 [13] Shwartz2015	[7] TR9856 [8] SemEval 2007 [9] All-Words [14] Zeichner2012	[3] SCWS
		context	

Figure 1: A map of prominent lexical inference datasets. **Word similarity**: [1] Zesch et al. (2008), [2] Hill et al. (2014), [3] Huang et al. (2012), [4] Wieting et al. (2015). **Term relatedness**: [5] Zesch et al. (2008), [6] Bruni et al. (2014), [7] Levy et al. (2015). **Lexical substitution**: [8] McCarthy and Navigli (2007), [9] Kremer et al. (2014), **Lexical inference**: [10] Kotlerman et al. (2010), [11] Turney and Mohammad (2014), [12] Levy et al. (2014), [13] Shwartz et al. (2015), [14] Zeichner et al. (2012), [15] Pavlick et al. (2015) (see 2.2), [16].

$\equiv$	Equivalence	is the same as
$\sqsubset$	Forward Entailment	is more specific than
$\sqsupset$	Reverse Entailment	is more general than
ˆ	Negation	is the exact opposite of
|	Alternation	is mutually exclusive with
$\sim$	Other-Related	is related in some other way to
#	Independence	is not related to

Table 2: Semantic relations in PPDB 2.0. Like Pavlick et al., we conflate *negation* and *alternation* into one relation.

2.2 PPDB with Semantic Relations

The PPDB paraphrase database (Ganitkevitch et al., 2013) is a huge resource of automatically derived paraphrases. In recent years, it has been used for quite many semantic tasks, such as semantic parsing (Wang et al., 2015), recognizing textual entailment (Noh et al., 2015), and monolingual alignment (Sultan et al., 2014).

Recently, as part of the PPDB 2.0 release, Pavlick et al. (2015) re-annotated PPDB with fine-grained semantic relations, following natural logic (MacCartney and Manning, 2007) (see table 2). This was done by first annotating a subset of PPDB pharaphase-pairs that appeared in the SICK dataset of textual entailment (Marelli et al., 2014). Annotators were instructed to select the appropriate semantic relation that holds for each paraphrase pair. These human annotations were later used to train a classifier and predict the semantic relation for all paraphrase pairs in PPDB. Considering the widespread usage of PPDB in applications, this extension may likely lead to applying lexical inferences based on such fine-grained semantic relations.

In this paper, we focus on human-annotated datasets, and therefore find the above mentioned subset of human-annotated paraphrases particularly relevant; we refer to this dataset as `PPDB-fine-human`. This dataset, as well as the PPDB 2.0 automatically created resource, are still missing a key feature in lexical inference, since the semantic relation for each paraphrase pair is specified out of context.

3 Dataset Construction Methodology

In this section, we present a methodology of adding context to lexical inference datasets, that we apply on `PPDB-fine-human`.

3.1 Selecting Phrase-Pairs

`PPDB-fine-human` is a quite large dataset (14k pairs), albeit with some phrase-pairs that are less useful for our purpose. We therefore applied the following filtering and editing on the phrase pairs:

Relation Types We expected that phrase pairs that were annotated out-of-context as *independent* will remain independent in almost every context; indeed, out of a sample of 100 such pairs that we annotated within context, only 8% were annotated with another semantic relation. As this was too sparse to justify the cost of human annotations, we chose to omit such phrase pairs.

Grammaticality-based Filtering Many phrases in `PPDB-fine-human` are ungrammatical, e.g. *boy is*. We consider such phrases less useful for our purpose, as semantic applications usually apply lexical inferences on syntactically coherent constituents. We therefore parse the original SICK (Marelli et al., 2014) sentences containing these phrases, and omit pairs in which one of the phrases is not a constituent.

Filtering Trivial Pairs In order to avoid trivial paraphrase pairs, we filter out inflections *(Iraq, Iraqi)* and alternate spellings *(center, centre)*, by omitting pairs that share the same lemma, or those that have Levenshtein distance ≤ 3. In addition, we omit pairs that have lexical overlaps *(a young lady, lady)* and filter out pairs in which one of the two phrases is just a stop word.

Removing Determiners The annotation seems to be indifferent to the presence of a determiner, e.g., the labelers annotated all of *(kid, the boy)*, *(the boy, the kid)*, and *(a kid, the boy)* as *reverse entailment*. To avoid repetitive pairs, and to get a single "normalized" phrase, we remove preceding determiners, e.g., yielding *(kid, boy)*.

Finally, it is interesting to note that `PPDB-fine-human` includes term-pairs in which terms are of different grammatical categories. Our view is that such cross-category term-pairs are often relevant for semantic inference (e.g. *(bicycle, riding)*) and therefore we decided to stick to the PPDB setting, and kept such pairs.

At the end of this filtering process we remained with 1385 phrase pairs from which we sampled 375 phrase pairs for our dataset, preserving the relative frequency across relation types in PPDB.

3.2 Adding Context Sentences

We used Wikinews[2] to extract context sentences. We used the Wikinews dump from November 2015, converted the Wiki Markup to clean text using WikiExtractor[3], and parsed the corpus using spaCy.[4]

For each (x, y) phrase-pair, we randomly sampled 10 sentence-pairs of the form (s_x, s_y), such that s_x contains x and s_y contains y. In the sampling process we require, for each of the two terms, that its 10 sentences are taken from different Wikinews articles, to obtain a broader range of the term's senses. This yields 10 tuples of the form (x, y, s_x, s_y) for each phrase pair and 3750 tuples in total.[5]

We split the dataset to 70% train, 25% test, and 5% validation sets. Each of the sets contains different term-pairs, to avoid overfitting for the most common relation of a term-pair in the training set.

3.3 Annotation Task

Our annotation task, carried out on Amazon Mechanical Turk, followed that of Pavlick et al. (2015). We used their guidelines, and altered them only to consider the contexts. We instructed annotators to select the relation that holds between the terms (x and y) while interpreting each term's meaning *within* its given context (s_x and s_y). To ensure the quality of workers, we applied a qualification test and required a US location, and a 99% approval rate for at least 1,000 prior HITS. We assigned each annotation to 5 workers, and, following Pavlick et al. (2015), selected the gold label using the majority rule, breaking ties at random. We note that for 91% of the examples, at least 3 of the annotators agreed.[6]

The annotations yielded moderate levels of agreement, with Fleiss' Kappa $\kappa = 0.51$ (Landis and Koch, 1977). For a fair comparison, we replicated the original out-of-context annotation on a sample of 100 pairs from our dataset, yielding agreement of $\kappa = 0.46$, while the in-context agreement for these pairs was $\kappa = 0.51$. As expected, adding context improves the agreement, by directing workers toward the same term senses while revealing rare senses that some workers may miss without context.[7]

4 Analysis

Figure 2 displays the confusion matrix of relation annotations in context compared to the out-of-context annotations. Most prominently, while the original relation holds in many of the contexts, it is also common for term-pairs to become independent. In some cases, the semantic relation is changed (as in table 1).

[2] `https://en.wikinews.org/`

[3] `https://github.com/attardi/wikiextractor`

[4] `http://spacy.io/`

[5] Our dataset is comparable in size to most of the datasets in Figure 1. In particular, the SCWS dataset (Huang et al., 2012), which is the most similar to ours, contains 2003 term-pairs with context sentences.

[6] We also released an additional version of the dataset, including only the agreeable 91%.

[7] The gap between the reported agreement in Pavlick et al. (2015) ($\kappa = 0.56$) and our agreement for out-of-context annotations ($\kappa = 0.46$) may be explained by our filtering process, removing obvious and hence easily consensual pairs.

		in-context					
		≡	⊏	⊐	|	∼	#
out-of-context	≡	60.54	6.9	0.38	0.38	9.58	22.22
	⊏	2.96	41.13	1.41	0	11.69	42.82
	⊐	5.97	1.67	37.92	2.08	13.19	39.17
	|	1.25	1.88	5.42	41.46	2.92	47.08
	∼	1.52	0.7	2.03	4.56	31.46	59.75

Figure 2: percentages of each relation annotation in-context, for annotations out-of-context. The diagonal shows out-of-context relations that hold in-context, and the last column shows term-pairs that become independent, usually due to sense-shifts. In all other cells, semantic relations are changed. Recall that we didn't annotate out-of-context independent pairs.

	precision	recall	F_1
PPDB-fine-human	**0.722**	0.380	0.288
PPDB2 classifier	0.611	0.565	0.556
in-context classifier	0.677	**0.685**	**0.670**

Table 3: Baseline performance on the test set (mean over all classes). (1) `PPDB-fine-human` manual annotations (out-of-context). (2) PPDB 2.0 classifier predictions (out-of-context). (3) our context-sensitive logistic regression classifier. Like Pavlick et al., we conflate the *forward entailment* and *reverse entailment* relations in all baselines.

4.1 Baseline Results

To demonstrate our dataset's utility, we report several baseline performances on our test set (table 3). The first two are context-insensitive, assigning the same label to a term-pair in all its contexts; the first assigns manual labels from `PPDB-fine-human`, and the second assigns PPDB 2.0 classifier predictions. We also trained a context-sensitive logistic regression classifier on our train set, using the available PPDB 2.0 features, plus additional context-sensitive features. To represent words as vectors, we used pretrained GloVe embeddings of 300 dimensions, trained on Wikipedia (Pennington et al., 2014), and added the following features:

$$max_{w \in s_y} \vec{x} \cdot \vec{w} \qquad (1)$$

$$max_{w \in s_x} \vec{y} \cdot \vec{w} \qquad (2)$$

$$max_{w_x \in s_x, w_y \in s_y} \vec{w_x} \cdot \vec{w_y} \qquad (3)$$

(1) and (2) measure similarities between a term and its most similar term in the other term's context, and (3) measures the maximal word similarity across the contexts.

This context-sensitive method, trained on our dataset, notably outperforms context insensitive baselines, thus illustrating the potential utility of our dataset for developing fine-grained context-sensitive lexical inference methods. Yet, the absolute performance is still mediocre, emphasizing the need to develop better such methods, using our dataset or similar ones created by our methodology.

5 Conclusion

In this paper, we presented a methodology for adding context to context-insensitive lexical inference datasets, and demonstrated it by creating such dataset over PPDB 2.0 fine-grained paraphrase-pair annotations. We then demonstrated that our dataset can indeed be used for developing fine-grained context-sensitive lexical inference methods, which outperform the corresponding context-insensitive baselines.

Acknowledgments

We would like to thank Ellie Pavlick and Chris Callison-Burch for their assistance and insightful comments.

This work was partially supported by an Intel ICRI-CI grant, the Israel Science Foundation grant 880/12, and the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *JAIR*, 49:1–47.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *HLT-NAACL*, pages 758–764.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456.*

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL 2012*, pages 873–882. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us-analysis of an all-words lexical substitution corpus. In *EACL 2014*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Omer Levy, Ido Dagan, and Jacob Goldberger. 2014. Focused entailment graphs for open ie propositions. In *CoNLL 2014*, pages 87–97, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ran Levy, Liat Ein-Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. 2015. Tr9856: A multi-word term relatedness benchmark. In *ACL 2015*, page 419.

Bill MacCartney and Christopher D Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200. Association for Computational Linguistics.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Tae-Gil Noh, Sebastian Padó, Vered Shwartz, Ido Dagan, Vivi Nastase, Kathrin Eichler, Lili Kotlerman, and Meni Adler. 2015. Multi-level alignments as an extensible representation basis for textual entailment algorithms. **SEM 2015*, page 193.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard H Hovy. 2007. Isp: Learning inferential selectional preferences. In *HLT-NAACL*, pages 564–571.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *ACL 2015*, Beijing, China, July. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543.

Vered Shwartz, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. Learning to exploit structured resources for lexical inference. *CoNLL 2015*, page 175.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *TACL 2014*, 2:219–230.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *ACL 2007*, page 456.

Peter D Turney and Saif M Mohammad. 2014. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(03):437–476.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *ACL 2015*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *TACL 2015*, 3:345–358.

Naomi Zeichner, Jonathan Berant, and Ido Dagan. 2012. Crowdsourcing inference-rule evaluation. In *ACL 2012*, pages 156–160. Association for Computational Linguistics.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, pages 861–866.