Adding Semantics to Data-Driven Paraphrasing

Ellie Pavlick[1] Johan Bos[2] Malvina Nissim[2] Charley Beller[3] Benjamin Van Durme[4] Chris Callison-Burch[1]
[1]Computer and Information Science Department, University of Pennsylvania
[2]Center for Language and Cognition Groningen, University of Groningen
[4]Human Language Technology Center of Excellence, Johns Hopkins University
[3]IBM Watson Group

Abstract

We add an interpretable semantics to the paraphrase database (PPDB). To date, the relationship between phrase pairs in the database has been weakly defined as approximately equivalent. We show that these pairs represent a variety of relations, including directed entailment (*little girl/girl*) and exclusion (*nobody/someone*). We automatically assign semantic entailment relations to entries in PPDB using features derived from past work on discovering inference rules from text and semantic taxonomy induction. We demonstrate that our model assigns these relations with high accuracy. In a downstream RTE task, our labels rival relations from WordNet and improve the coverage of a proof-based RTE system by 17%.

Riots in Denmark were sparked by 12 editorial cartoons that were offensive to Muhammad.

12	$\equiv$	Twelve
editorial cartoons	$\sqsupset$	illustrations
offensive	$\sqsupset$	insulting
Muhammad	$\equiv$	the prophet
sparked	$\sqsupset$	caused
riots	$\sqsupset$	unrest
in Denmark	$\mid$	in Jordan

Twelve illustrations insulting the prophet caused unrest in Jordan.

Figure 1: An example sentence pair for the RTE task. In order for a system to conclude that the premise (top) does not entail the hypothesis (bottom), it should recognize that *sparked* implies *caused* but that *in Denmark* precludes *in Jordan*. These phrase-level entailment relationships are modeled by natural logic.

1 Motivation

A basic precursor to language understanding is the ability to recognize when two expressions mean the same thing. Different expressions of the same information is the central problem addressed by paraphrasing and the closely related task of recognizing textual entailment (RTE). In RTE, a system is given two pieces of text, often called the *text* (T) and the *hypothesis* (H), and asked to determine whether T entails H, T contradicts H, or T and H are unrelatable (Figure 1). In contrast, data-driving paraphrasing typically sidesteps developing a clear definition of "meaning the same thing" and instead "assume[s] paraphrasing is a coherent notion and concentrate[s] on devices that can produce paraphrases" (Barzilay, 2003). Recent work on paraphrase extraction has resulted in enormous paraphrase collections (Lin and Pantel, 2001; Dolan et al., 2004; Ganitkevitch et al., 2013), but the usefulness of these collections is limited by the fast-and-loose treatment of the meaning of paraphrases. One concrete definition that is sometimes used for paraphrases requires that they be bidirectionally entailing (Androutsopoulos and Malakasiotis, 2010). That is, in terms of RTE, it is assumed that if P is a paraphrase of Q, then P entails Q and Q entails P. In reality, paraphrases are often more nuanced (Bhagat and Hovy, 2013), and the entries in most paraphrase resources certainly do not match this definition. For instance, Lin and Pantel (2001) extracted 12 million "inference rules" from monolingual text by exploiting shared dependency contexts. Their method learns paraphrases that are truly meaning equivalent, but it just as readily learns contradictory pairs such as $\langle X\,rises, X\,falls\rangle$. Ganitkevitch et al. (2013) extract over 150 million paraphrase rules by pivoting through foreign translations. This bilingual method often learns hypernym/hyponym pairs, e.g. due to variation in the discourse structure of translations (Callison-

Equivalent	Entailment	Exclusion	Other relation	Unrelated
look at/watch	little girl/girl	close/open	swim/water	girl/play
a person/someone	kuwait/country	minimal/significant	husband/marry to	found/party
clean/cleanse	tower/building	boy/young girl	oil/oil price	profit/year
away/out	the cia/agency	nobody/someone	country/patriotic	man/talk
distant/remote	sneaker/footwear	blue/green	drive/vehicle	car/family
the phone/the telephone	heroin/drug	france/germany	family/home	holiday/series
last autumn/last fall	doe/deer	least three/least two	basketball/court	green/tennis
illegal entry/smuggling	typhoon/storm	child/mother	playing/toy	sunday/tour
approve/to ratify	seriously injure/injure	in front/on the side	islamic/jihad	city/south
alliance of/coalition between	sunglasses/glasses	oppose/support	delay/time	back/view

Table 1: Examples of different types of entailment relations appearing in PPDB.

Burch, 2007), and unrelated pairs, e.g. due to misalignments or polysemy in the foreign language.

The unclear semantics severely limits the applicability of paraphrase resources to natural language understanding (NLU) tasks. Some efforts have been made to identify directionality of paraphrases (Bhagat et al., 2007; Kotlerman et al., 2010), but tasks like RTE require even richer semantic information. For example, in the T/H pair shown in Figure 1, a system needs information not only about equivalent words (*12/twelve*) and asymmetric entailments (*riots/unrest*), but also semantic exclusion (*Denmark/Jordan*). Such lexical entailment relations are captured by *natural logic*, a formalism which views natural language itself as a meaning representation, eschewing external representations such as First Order Logic (FOL). This is a great fit for automatically extracted paraphrases, since the phrase pairs themselves can be used as the semantic representation with minimal additional annotation. But as is, paraphrase resources lack such annotation.

As a result, NLU systems rely on manually built resources like WordNet, which are limited in coverage and often lead to incorrect inferences (Kaplan and Schubert, 2001). In fact, in the most recent RTE challenge, over half of the submitted systems used WordNet (Pontiki et al., 2014). Even the NatLog system (MacCartney and Manning, 2007), which popularized natural logic for RTE, relied on WordNet and did not solve the problem of assigning natural logic relations at scale.

The main contributions of this paper are:

- We add a concrete, interpretable semantics to the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013), the largest paraphrase resource currently available. We give each entry in the database a label describing the entailment relationship between the phrases.
- We develop a statistical model to predict these relations. The enormous size of PPDB– over 77 million phrase pairs!– makes it impossible to perform this task manually. Our wide range of monolingual and bilingual features results in high intrinsic accuracy.
- We demonstrate improvements to a proof-based RTE system, showing that our automatic labels increase the number of proofs that it is able to find by 17%, while maintaining the same accuracy as when using gold-standard, manual labels.

2 Related Work

Lexical entailment resources Approaches to paraphrase identification have exploited signal from distributional contexts (Lin and Pantel, 2001; Szpektor et al., 2004), comparable corpora (Dolan et al., 2004; Xu et al., 2014), and graph structures (Berant et al., 2011; Brockett et al., 2013). These approaches are scalable, but they often assume that all relations are equivalence relations (Madnani and Dorr, 2010). Several efforts have attempted to build or augment lexical ontologies automatically, to discover other types of lexical relations like hypernyms. Most of these approaches rely on lexico-syntactic patterns. Hearst (1992) searched for hand-written patterns (e.g. "an X is a Y") in a large corpus in order to learn taxonomic relations between nouns. Snow et al. (2006) used dependency parses to automatically learn such patterns, which they used to augment WordNet with new hypernym relations. Similar monolingual signals have been used to learn fine-grained relationships between verbs, such as *enablement* and *happens-before* (Chklovski and Pantel, 2004; Hashimoto et al., 2009).

Recognizing Textual Entailment The shared RTE tasks (Dagan et al., 2006) have been a springboard for research in natural language inference,

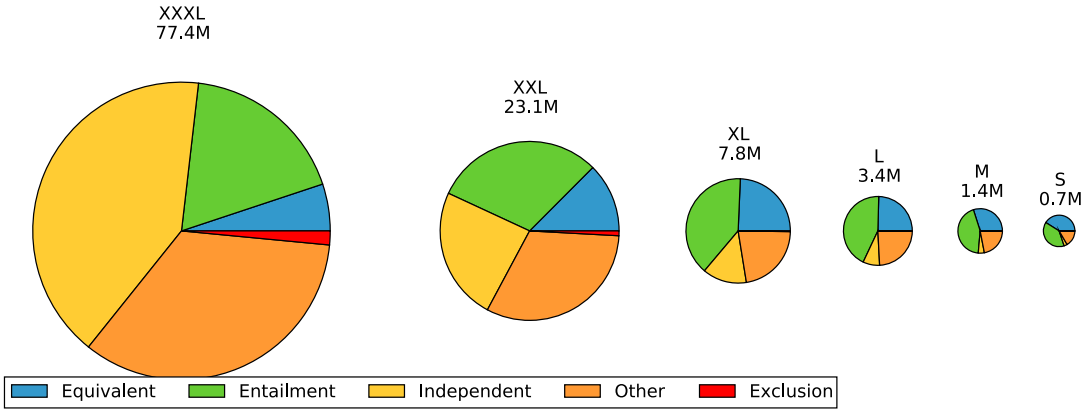

Figure 2: Distribution of entailment relations in different sizes of PPDB. Distributions are estimated from our manual annotations of randomly sampled pairs. PPDB-XXXL contains over 77MM paraphrase pairs (where the majority type is independent), compared to only 700K in PPDB-S (where the majority type is equivalent).

using data motivated by the applications to information retrieval, information extraction, summarization, machine translation evaluation, and more recently, question answering (Giampiccolo et al., 2007) and essay grading (Clark et al., 2013). RTE systems vary considerably in their choice of representation and inference procedure. In the most recent shared task on RTE, some systems used deep logical representations of text, allowing them to invoke theorem provers (Bjerva et al., 2014) or Markov Logic Networks (Beltagy et al., 2014) to perform the inference, while others used shallower representations, relying on machine learning to perform inference (Lai and Hockenmaier, 2014; Zhao et al., 2014). Systems based on natural logic (MacCartney and Manning, 2007) use natural language as a representation, but still perform inference using a structured algebra rather than a statistical model. Regardless of the inference procedure, improvements to external lexical resources can improve RTE systems across the board (Clark et al., 2007).

3 The Paraphrase Database (PPDB)

PPDB is currently the largest available collection of paraphrases. Compared to other paraphrase resources such as the DIRT database (12 million rules) (Lin and Pantel, 2001) and the MSR paraphrase phrase table (13 million) (Dolan et al., 2004), PPDB contains over 150 million paraphrase rules covering three paraphrase types– lexical (single word), phrasal (multiword), and syntactic restructuring rules. We focus on lexical and phrasal paraphrases, of which there are over 77 million rules. Of these, a large fraction are true paraphrases– either equivalent (*distant/remote*) or asymmetric entailment (*girl/little girl*)– but many are not. PPDB contains some pairs which are related by semantic exclusion (*nobody/someone*), some of which are related by something other than entailment (*swim/water*), and some which are simply unrelated (*car/family*). Table 1 gives examples of pairs in PPDB falling into each of these categories.

PPDB is released in six sizes (S, M, L, XL, XXL and XXXL), which fall roughly on a continuum from highest precision and lowest recall to lowest average precision and highest recall. Figure 2 shows how the distribution of entailment relations differs across the sizes of PPDB.[1] Our goal is to make these relations explicit, by providing annotations for each phrase pair. Because of the enormous scale of PPDB, this annotation must be done automatically.

4 Selection of Paraphrases

In this paper we focus on paraphrases pairs from PPDB that occur in RTE data. We use the recent SICK dataset from in the 2014 SemEval RTE challenge (Marelli et al., 2014) for our experiments. The data consists of 10K sentences split roughly evenly into training and testing sets. The sentence pairs are labeled using a 3-way entailment classification: ENTAILMENT, (29%) CONTRADICTION (15%), or NEUTRAL (56%). We consider all phrase pairs from PPDB $\langle p_1, p_2 \rangle$ up to three words in length such that there is some T/H sentence pair in which p_1 appears in T and p_2 appears

[1]These distributions were estimated based on a random sample of pairs drawn from each size of PPDB, annotated on MTurk as described in Section 5

Lexical	We use the lemmas, POS tags, and phrase lengths of p_1 and p_2, the substrings shared by p_1 and p_2, and the Levenstein, Jaccard, and Hamming distances between p_1 and p_2.
Distributional	Given a dependency context vectors for p_1 and p_2, we compute the number of shared contexts, and the Jaccard, Cosine, Lin1998, Weeds2004, Clarke2009, and Szpektor2008 similarities between the vectors.
Paraphrase	We include 33 paraphrase features distributed with PPDB, which include the paraphrase probabilities as computed in Bannard and Callison-Burch (2005). We refer the reader to Ganitkevitch and Callison-Burch (2014) for a complete description of all of the features included with PPDB.
Translation	We include the number of foreign language "pivots" (translations) shared by p_1 and p_2 for each of 24 languages used in the construction of PPDB, as a fraction of the total number of translations observed for each of p_1 and p_2.
Path	We include a sparse vector of all lexico-syntactic patterns (paths through a dependency parse) which are observed between p_1 and p_2 in the Annotated Gigaword corpus (Napoles et al., 2012).
WordNet	We include binary features indicating whether WordNet classifies p_1 and p_2 according to any of the following relations: synonym, hypernym, hyponym, antonym, holonym, meronym, cause, entailment, derivationally-related, similar-to, also-see, or attribute.

Figure 3: Summary of features extracted for each phrase pair $\langle p_1, p_2 \rangle$. Full descriptions of the features used are given in the supplementary material.

in H. Roughly 55% of the word types and 5% of the phrase (bigram and trigram) types in the SICK data appear in PPDB. This gives us a list of 9,600 pairs, half from the training sentences, which we use for development in Section 6, and half from the test sentences, which we use for evaluation in Section 7.

The SICK data has a relatively small vocabulary, with 86% of words types and <1% of the phrase types covered by WordNet. Still, over half of the words in SICK which are covered by PPDB do not appear in WordNet. In general, PPDB covers a much larger vocabulary (1.6MM words) than does WordNet (155K words), and we expect the potential benefit of using PPDB in addition to or in place of WordNet to be larger on datasets with richer vocabularies.

5 Entailment Relations

We use the relations from Bill MacCartney's thesis on natural language inference as the basis for our categorization of relations (MacCartney, 2009). He outlines 7 basic entailment relationships:[2]

Equivalence (P≡Q): $\forall x[P(x) \leftrightarrow Q(x)]$
Forward Entailment (P⊏Q): $\forall x[P(x) \rightarrow Q(x)]$
Reverse Entailment (P⊐Q): $\forall x[Q(x) \rightarrow P(x)]$
Negation (P^Q): $\forall x\ [P(x) \leftrightarrow \neg\ Q(x)]$
Alternation (P|Q): $\forall x\ \neg[P(x) \wedge Q(x)]$
Cover (P⌣Q): $\forall x[P(x) \vee Q(x)]$
Independence (P#Q): All other cases.

[2]To further clarify the definitions here: "negation" is XOR (exclusive disjunction), "alternation" is NAND, and "cover" is OR (inclusive disjunction)

These relations are based on the theory of *natural logic*, meaning they are defined between pairs of natural language expressions rather than requiring an external formal representation. This makes them an ideal fit for the phrase pairs in in PPDB and similar automatically-constructed paraphrase resources.

Nat. Log.	This work	MTurk description
≡	≡	X is the same as Y
⊏	⊏	X is more specific than/is a type of Y
⊐	⊐	X is more general than/encompasses Y
^	¬	X is the opposite of Y
|		X is mutually exclusive with Y
#	∼	X is related in some other way to Y
	#	X is not related to Y

Table 2: Column 1 gives the semantics of each label under MacCartney's Natural Logic. Column 2 gives the notation we use throughout the remainder of this paper. Column 3 gives the description that was shown to Turkers.

Annotation We use Amazon Mechanical Turk (MTurk) to collect labels for our phrase pairs. We asked workers to choose between the options show in Table 2, which represent a modified version of MacCartney's relations. We replace negation (^) with the weaker notion of "opposites," effectively merging it with the alternation (|) relation; we split the independent (#) class into two cases: truly independent phrases and phrases which are related by something other than entailment (which we denote ∼). We omit the cover (⌣) relation entirely, as its practicality is not obvious. We show each pair to 5 workers, taking the majority label as truth. Each HIT consisted of two control questions taken from WordNet. Workers achieved good accuracies on our controls (82% overall) and moder-

Cosine Similarity		Monolingual (symmetric)		Monolingual (asymmetric)		Bilingual	
$\sqsupset$	shades/the shade	$\neg$	large/small	$\sqsupset$	boy/little boy	$\equiv$	dad/father
$\sqsupset$	yard/backyard	$\equiv$	few/several	$\sqsupset$	man/two men	$\sqsupset$	some kid/child
$\#$	each other/man	$\neg$	different/same	$\sqsupset$	child/three children	$\equiv$	a lot of/many
$\sqsupset$	picture/drawing	$\neg$	other/same	$\equiv$	is playing/play	$\equiv$	female/woman
$\sim$	practice/target	$\neg$	put/take	$\sqsupset$	side/both sides	$\equiv$	male/man

Table 3: Top scoring pairs (x/y) according to various similarity measures, along with their manually classified entailment labels. Column 1 is cosine similarity based on dependency contexts. Column 2 is based on Lin (1998), column 3 on Weeds (2004), and column 4 is a novel feature. Precise definitions of each metric are given in the supplementary material.

ate levels of agreement (Fleiss's $\kappa = 0.56$) (Landis and Koch, 1977). For a fuller discussion of the annotation, refer to the supplementary material.

6 Automatic Classification

We aim to build a classifier to automatically assign entailment types to entries in the PPDB, and to demonstrate that it performs well both intrinsically and extrinsically. We fix the direction of the $\sqsubset$ and $\sqsupset$ relations to create a single class and train a logistic regression classifier to distinguish between the 5 classes $\{\#, \equiv, \sqsupset, \neg, \sim\}$. We compute variety of basic lexical features and WordNet features (summarized in Figure 3). We categorize the remaining features into two broad groups: monolingual features, which are based on observed usage in the Annotated Gigaword corpus (Napoles et al., 2012), and bilingual features, which are based on translation probabilities observed in bilingual parallel corpora. Full descriptions of all the features used are provided in the supplementary material.

6.1 Monolingual features

Path features Snow et al. (2004) used lexico-syntactic patterns to mine taxonomic relations (hypernyms and hyponyms) between noun pairs. They were able to verify the earlier work of Hearst (1992) which found that certain patterns, e.g. *X and other Y*, are strong indicators of hypernymy. Using similar path features, we learn new patterns to differentiate between more subtle relations. For example, we learn the pattern *separate X from Y* is highly indicative of the $\neg$ relation. We learn that the pattern *X including Y* suggests $\sqsupset$ more than it suggests $\equiv$ whereas the pattern *X known as Y* suggests $\equiv$ more than $\sqsupset$. Table 4 gives examples of some of the paths most indicative of the $\neg$ relation.

Distributional features Lin and Pantel (2001) attempted to mine inference rules from text by finding paths in a dependency tree which connect the same nouns. The intuition is that good paraphrases should tend to modify and be modified by the same words. Given context vectors, Lin and Pantel (2001) used a symmetric similarity metric (Lin, 1998) to find candidate paraphrases. We build dependency context vectors for each word in our data and compute both symmetric as well as more recently proposed asymmetric similarity measures (Weeds et al., 2004; Szpektor and Dagan, 2008; Clarke, 2009), which are potentially better suited for identifying $\sqsupset$ paraphrases. Table 3 gives a comparison of the pairs which are considered "most similar" according to several of these metrics.

in X and in Y	*in foods and in beverages*
separate X from Y	*separate the old from the young*
to X and/or to Y	*to the left or to the right*
from X to Y	*from 7 a.m. to 10 p.m.*
more/less X than Y	*more harm than good*

Table 4: Top paths associated with the $\neg$ class.

6.2 Bilingual features

We explore a variety of bilingual features, which we expect to provide complimentary signal to the monolingual features. Each pair in PPDB is associated with several paraphrase probabilities, which are based on the probabilities of aligning each word to the foreign "pivot" phrase (a foreign translation shared by the two phrases), computed as described in Bannard and Callison-Burch (2005). We also compute the total number of shared foreign translations for each phrase pair. Table 3 shows the highest ranked pairs by this bilingual similarity score, in comparison to several of the monolingual scores.

6.3 Analysis

Table 5 shows an ablation analysis. The bilingual features are especially important for distinguishing the $\equiv$ class, and the path and WordNet features are important for the $\neg$ class. The lexical features show strong performance across the board; this is often because they capture negation words (e.g. *no*) and substring features (*little boy* $\sqsubset$ *boy*).

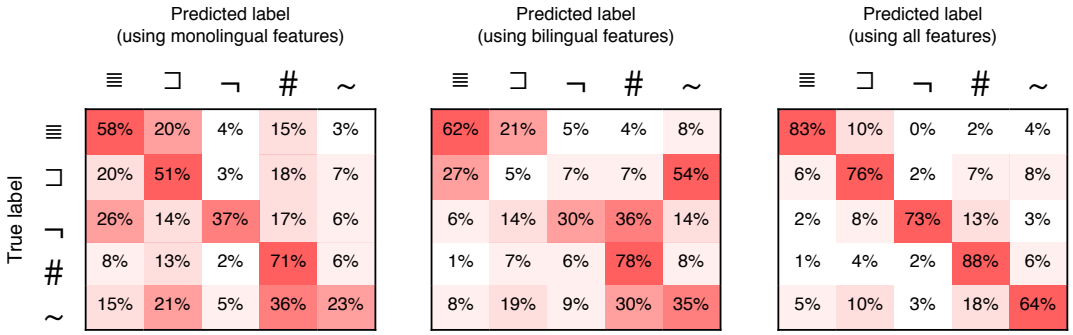

Figure 4: Confusion matrices for classifier trained using only monolingual features (distributional and path) versus bilingual features (paraphrase and translation). True labels are shown along rows, predicted along columns. The matrix is normalized along rows, so that the predictions for each (true) class sum to 100%. The confusion matrices reflect classifier's performance on held-out phrase pairs from the SICK test set.

		Δ F1 when excluding					
	All	Lex.	Dist.	Path	Para.	Tran.	WN
#	79	-2.0	-0.2	-1.2	-1.7	-0.2	-0.1
≡	57	-3.5	+0.2	-0.7	-2.4	-3.7	+0.5
⊐	68	-4.6	-0.3	-0.8	-0.8	-0.7	-1.6
¬	49	-4.0	-0.8	-2.9	+0.3	-0.0	-2.2
∼	51	-4.9	-0.5	-0.7	-1.2	-0.9	-0.3

Table 5: F1 measure (×100) achieved by entailment classifier using 10-fold cross validation on the training data.

Table 3 shines some light onto the differences between monolingual and bilingual similarities. While the monolingual asymmetric metrics are good for identifying ⊐ pairs, the symmetric metrics consistently identify ¬ pairs; none of the monolingual scores we explored were effective in making the subtle distinction between ≡ pairs and the other types of paraphrase. In contrast, the bilingual similarity metric is fairly precise for identifying ≡ pairs, but provides less information for distinguishing between types of non-equivalent paraphrase. These differences are further exhibited in the confusion matrices shown in Figure 4; when the classifier is trained using only monolingual features, it misclassifies 26% of ¬ pairs as ≡, whereas the bilingual features make this error only 6% of the time. On the other hand, the bilingual features completely fail to predict the ⊐ class, calling over 80% of such pairs ≡ or ∼.

7 Evaluation

7.1 Intrinsic Evaluation

We test the performance of our classifier intrinsically, through its ability to reproduce the human labels for the phrase pairs from the SICK test sentences. Table 7 shows the precision and recall achieved by the classifier for each of our 5 entailment classes. The classifier is able to achieve an overall 79% accuracy, reaching >70% precision while maintaining good levels of recall on all classes.

True	Pred.	N	Example misclassifications
∼	#	169	boy/little, an empy/the air
#	∼	114	little/toy, color/hair
⊐	∼	108	drink/juice, ocean/surf
⊐	#	97	in front of/the face of, vehicle/horse
⊐	≡	83	cat/kitten, pavement/sidewalk
≡	⊐	46	big/grand, a girl/a young lady
⊐	¬	29	kid/teenager, no small/a large
¬	⊐	29	old man/young man, a car/a window
#	≡	15	a person/one, a crowd/a large
≡	#	9	he is/man is, photo/still
≡	¬	1	girl is/she is

Table 6: Example misclassifications from some of the most frequent and most interesting error categories.

Figure 4 shows the classifier's confusion matrix and Table 6 shows some examples of common and interesting error cases. The majority of errors (26%) come from confusing the ∼ class with the # class. This mistake is not too concerning from an RTE perspective since ∼ can be treated as a special case of # (Section 5). There are very few cases in which the classifier makes extreme errors, e.g. confusing ≡ with ¬ or with #; some interesting examples of such errors arise when the phrases contain pronouns (e.g. *girl* ≡ *she*) or when the relation uses a highly infrequent word sense (e.g. *photo* ≡ *still*).

7.2 The Nutcracker RTE System

To further test our classifier, we evaluate the usefulness of the automatic entailment predictions in a downstream RTE task. We run our experiments using Nutcracker, a state-of-the-art RTE system based on formal semantics (Bjerva et al., 2014).

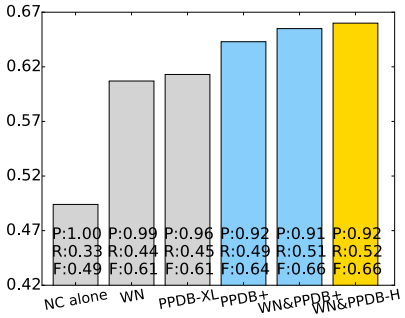

Figure 5: ENTAILMENT

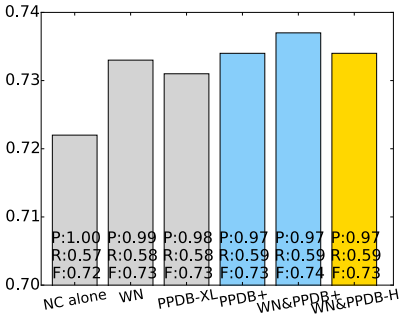

Figure 6: CONTRADICTION

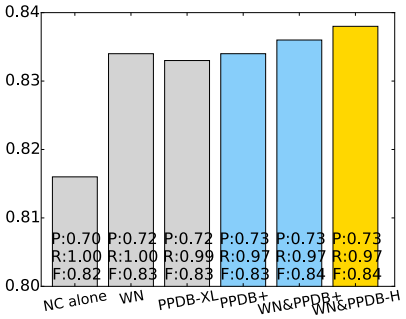

Figure 7: NEUTRAL

Figure 8: F1 measures achieved by Nutcracker on SICK test data when using various KBs. Baselines are in gray, this work in blue, human references in gold. PPDB-XL refers to a run in which every pair which appears in PPDB is assumed to be equivalent. PPDB-H refers to a run in which manual labels were used to generate axioms. PPDB+ refers to runs in which the automatic classifications were used to generate axioms. In some cases, better proof coverage causes NC to find incorrect proofs, illustrated by the decreased performance on CONTRADICTION when using PPDB-H. For example, using PPDB-H, NC finds an inconsistency for the pair *Someone is not playing piano./A person is playing a keyboard.* Using the PPDB+, in which *piano/keyboard* is falsely classified as #, NC fails to find a proof and so correctly guesses NEUTRAL.

	Freq.	Precision	Recall	F score
#	39%	84.22	87.55	85.85
≡	8%	70.36	83.07	76.19
⊐	26%	79.81	76.00	77.85
¬	7%	73.73	73.33	73.53
∼	19%	70.57	63.70	66.96

Table 7: F1 measure (×100) achieved by entailment classifier on the held out phrase pairs from the sentences in SICK test.

	Acc.	# Proofs	Coverage
MFC	56.4	0	0%
NC alone	74.3	878	17.8%
+ WN	77.5	1,051	21.3%
+ PPDB-XL	77.5	1,091	22.1%
+ PPDB+	78.0	1,197	24.3%
+ WN, PPDB+	**78.4**	**1,230**	**25.0%**
+ WN, PPDB-H	*78.6*	*1,232*	*25.0%*

Table 8: Nutcracker's overall system accuracy and proof coverage when using different sources of axioms. Coverage is measured as the percent of sentence pairs for which NC's theorem prover or model builder is able to find a complete logical proof of either entailment or contradiction. When NC fails to find either type of proof, it guesses the most frequent class, NEUTRAL. NC alone uses no axioms. PPDB+ refers to the axioms generated automatically using the classifier described in this paper. PPDB-H refers axioms generated using the human labels on which the classifier was trained.

In the SemEval 2014 RTE challenge, this system performed in the top 5 out of the more than 20 participating systems (Marelli et al., 2014).

Given a text/hypothesis (T/H) pair, Nutcracker (NC) uses the Boxer parser (Bos, 2008) to produce a formal semantic representation of both T and H, which it translates into standard first-order logic. The logical formulae are passed to an off-the-shelf theorem prover, which searches for a logical entailment, and to a model builder, which attempts to find a logical contradiction. By default, when the system fails to find a proof for either entailment or inconsistency, it predicts the most frequent class (in our case, NEUTRAL). Therefore, NC relies heavily on lexical entailment resources in order to improve the recall of the theorem prover and model builder.

Baselines The most frequent class baseline is achieved by labeling every sentence pair as NEUTRAL, and results in an accuracy of 56%. A stronger baseline is obtained by running NC alone, without any external axioms; in this case, words are only equivalent if they are lemma-identical.

As an additional baseline, we generate a "basic" PPDB-XL[3] knowledge base (KB), which consists exclusively of axioms expressing synonym relationships. I.e. for every pair of phrases $\langle p_1, p_2 \rangle$ in PPDB-XL, the PPDB-XL KB contains the equivalence axiom $syn(p_1, p_2)$. We also generate the WordNet (WN) KB, which is the default used by NC. This KB consists of axioms for all synonyms, antonyms, and hypernyms in WN, which generate syn, $isnota$, and isa axioms, respectively.

PPDB+ We convert our classifier's predictions into a set of axioms for NC. When our classifier predicts ≡ we generate an syn axiom, when it predicts ⊐ we generate an isa axiom, and when it predicts ¬ we generate an $isnota$ axiom. # and ∼ do not generate any axioms. To handle the directionality of the ⊐ relation, we run the classifier

[3]We generated basic KBs for all six sizes of PPDB, but XL performed best.

True	PPDB+	WN	Text/Hypothesis pair
ENTAIL.	ENTAIL.	NEUTRAL	A **bride** in a white dress is running/A **girl** in a white dress is running.
ENTAIL.	NEUTRAL	ENTAIL.	A **lemur** is biting a person's finger./An **animal** is biting a person's finger.
CONTRA.	CONTRA.	NEUTRAL	**Someone** is playing a piano./There is **no one** playing a piano.
CONTRA.	NEUTRAL	CONTRA.	There is **no man** pouring oil into a **pan**./**A man** is pouring oil into a **skillet**.

Table 9: Examples of T/H pairs for which the system's prediction differed when using PPDB+ vs. WN.

over every pair in both directions, and we choose whichever direction and relation receives the highest confidence score to be the final prediction. We refer to this set of automatically-predicted axioms as PPDB+.

To calibrate our improvements, we also generate a KB using the human labels collected from MTurk, which we refer to as PPDB-Human or PPDB-H.

Results Table 8 reports NC's overall prediction accuracy and the number of proofs found when using each of the described KBs. Figure 8 shows the performance in terms of the precision and recall achieved for each of the three entailment classes: ENTAILMENT, CONTRADICTION, and NEUTRAL. Table 9 provides some examples of T/H pairs on which predictions differed using the PPDB+ compared to the WN KB, and Figure 9 shows some illustrative misclassifications.

Our automatic labels result in a 4% improvement in accuracy over the baseline of using NC alone (Figure 8), and a 15 point improvement in F1 measure for the entailment class (Table 8). By all performance measures, PPDB+ also outperforms WordNet as a source of axioms for NC. Moreover, adding PPDB+ to WordNet gives a 17% relative increase in the number of proofs found compared to using WordNet alone (Table 8). These additional proofs lead NC to make a greater number of correct predictions for the "right reasons" (i.e. finding a proof/contradiction) rather than by lucky guessing (recall NC guesses the most frequent class when it cannot find a proof).

For comparison, we run the same experiments using a KB of oracle human labels in place of the predicted labels in PPDB+. Using PPDB+, NC comes very close to the performance achieved when using PPDB-Human, demonstrating that the automatically generated PPDB+ provides as much utility to the end-to-end system as does a gold-standard resource.

8 Data Release

Upon publication, we are releasing a new PPDB fully annotated with semantic relations. We are also releasing the set of 14K manually labeled phrase pairs occurring in RTE data, and our software for extracting features and running the classifier, so that researchers can apply our model to their own paraphrase collections. This will constitute the largest lexical entailment resources available, while also offering new fine-grained annotation necessary for challenging NLU tasks. An evaluation of the predicted relations appearing in the entire Paraphrase Database (not just those occurring in RTE data) is given in the supplementary material.

9 Conclusion

We argue that a significant failing of recent work on data-driven paraphrasing is the weak definition of paraphrases as being more-or-less equivalent. In this paper, we show how a clear concept of semantics can be applied to large-scale paraphrase resources. In particular, the entailment relations given by natural logic are a great fit for paraphrase resources, since natural logic operates on pairs of natural language expressions (like the entries in PPDB). By classifying paraphrase entries with entailment relations, we provide them with an interpretable semantics. Our classifier uses extensive feature sets to scale natural logic to the enormous number of phrase pairs in PPDB. We rigorously evaluate our model, demonstrating high accuracy on an intrinsic task. On an extrinsic RTE task, our model's predictions allow an RTE system to find 17% more proofs and achieve a higher overall accuracy than when using WordNet's manual relations. Our new release of PPDB, annotated with semantic entailments, will dramatically improve PPDB's utility for NLU tasks.

Acknowledgements This research was supported by the Allen Institute for Artificial Intelligence (AI2), the Human Language Technology Center of Excellence (HLTCOE), and by gifts from the Alfred P. Sloan Foundation, Google, and

	# 38%	≡ 8%	⊐ 26%	¬ 7%	∼ 18%
# 40%	1730 (clear,very) (exhibit,hold) (walk,woman)	9 (cover,front) (photo,still) (woman who,woman with)	97 (hand,male) (man,police) (mountain,side)	49 (drive,park) (female,man) (flag,ship)	169 (child,park) (crowded,many) (note,write)
≡ 10%	15 (a big,very) (a lot,long) (face a,front of)	368 (a small,the little) (away,out) (block,slab)	83 (a gun,a weapon) (a weapon,gun) (legs,leg)	9 (another man,one man) (bike,biking) (young girl,young woman)	48 (a child,kid in) (and hold,and take) (his arms,his hands)
⊐ 24%	82 (device,guy) (something,talk) (the man,the phone)	46 (a call,phone call) (a group,bunch of) (another man,man)	1004 (camera,webcam) (kid,other child) (kid,the daughter)	29 (a car,a window) (a female,a man) (arms,his hands)	97 (a lady,girl) (field,playing) (girl,the lady)
¬ 7%	35 (a ball,a man) (a boy,little) (number,woman)	1 (girl is,she is)	29 (a boy,a teenager) (a kid,daughter) (kid,little girl)	275 (cat,dog) (morning,night) (type,write)	33 (dog,owner) (ground,water) (hat,vest)
∼ 17%	114 (leg,soccer) (perform,run) (sail,water)	19 (chef,cook) (fight,match) (race,ride)	108 (cut,saw) (face,hair) (the kid,the little)	13 (a boat,sail) (dress,suit) (light,the dark)	609 (ice,rink) (snow,snowy) (study by,study the)

Figure 9: Confusion matrix for classifier (with all features) on SICK test set. True labels and their distribution are shown along the columns, predicted along the rows.

Facebook. This material is based in part on research sponsored by the NSF under grant IIS-1249516 and DARPA under agreement number FA8750-13-2-0017 (the DEFT program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

The authors would like to thank Peter Clark, Bill MacCartney, Patrick Pantel and the anonymous reviews for their thoughtful suggestions.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604.

Regina Barzilay. 2003. *Information fusion for multi-document summarization: paraphrasing and generation*. Ph.D. thesis, Columbia University.

Islam Beltagy, Stephen Roller, Gemma Boleda, Katrin Erk, and Raymond J Mooney. 2014. UTexas: Natural language semantics using distributional semantics and probabilistic logic. *SemEval 2014*, page 796.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 610–619.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39.

Rahul Bhagat, Patrick Pantel, Eduard H Hovy, and Marina Rey. 2007. Ledir: An unsupervised algorithm for learning directionality of inference rules. In *EMNLP-CoNLL*, pages 161–170. Citeseer.

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. *SemEval 2014*, page 642.

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

Christopher John Brockett, Stanley Kok, and Dengyong Zhou. 2013. Locating paraphrases through utilization of a multipartite graph, July 9. US Patent 8,484,016.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *EMNLP*, volume 2004, pages 33–40.

Peter Clark, William R. Murray, John Thompson, Phil Harrison, Jerry Hobbs, and Christiane Fellbaum. 2007. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 54–59.

Peter Clark, Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 112–119.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May. European Language Resources Association.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata, and Jun'ichi Kazama. 2009. Large-scale verb entailment acquisition from the web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1172–1181. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545.

Aaron N Kaplan and Lenhart K Schubert. 2001. Measuring and improving the quality of world knowledge extracted from wordnet. *University of Rochester, Rochester, NY*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A denotational and distributional approach to semantics. *SemEval 2014*, page 329.

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

Dekang Lin and Patrick Pantel. 2001. DIRT – Discovery of Inference Rules from Text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 193–200.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Citeseer.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36.

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*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100.

Maria Pontiki, Haris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of SemEval, Dublin, Ireland*.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*, volume 17, pages 1297–1304.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 801–808.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 849–856.

Idan Szpektor, Hristo Tanev, Dr Dagan, Bonaventura Coppola, et al. 2004. Scaling web-based acquisition of entailment relations.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, 2.

Jiang Zhao, Tian Tian Zhu, and Man Lan. 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. *SemEval 2014*, page 271.