PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification

Ellie Pavlick[1] Pushpendre Rastogi[2] Juri Ganitkevitch[2]
Benjamin Van Durme[2,3] Chris Callison-Burch[1]
[1]Computer and Information Science Department, University of Pennsylvania
[2]Center for Language and Speech Processing, Johns Hopkins Universityy
[3]Human Language Technology Center of Excellence, Johns Hopkins University

Abstract

We present a new release of the Paraphrase Database. PPDB 2.0 includes a discriminatively re-ranked set of paraphrases that achieve a higher correlation with human judgments than PPDB 1.0's heuristic rankings. Each paraphrase pair in the database now also includes fine-grained entailment relations, word embedding similarities, and style annotations.

1 Introduction

The Paraphrase Database (PPDB) is a collection of over 100 million paraphrases that was automatically constructed by Ganitkevitch et al. (2013). Although it is relatively new, it has been adopted by a large number of researchers, who have demonstrated that it is useful for a variety of natural language processing tasks. It has been used for recognizing textual entailment (Beltagy et al., 2014; Bjerva et al., 2014), measuring the semantic similarity of texts (Han et al., 2013; Ji and Eisenstein, 2013; Sultan et al., 2014b), monolingual alignment (Yao et al., 2013; Sultan et al., 2014a), natural language generation (Ganitkevitch et al., 2011), and improved lexical embeddings (Yu and Dredze, 2014; Rastogi et al., 2015; Faruqui et al., 2015).

For any given input phrase to PPDB, there are often dozens or hundreds of possible paraphrases. There are several interesting research questions that arise because of the number and variety of paraphrases in PPDB. How can we distinguish between correct and incorrect paraphrases? Within the paraphrase sets, are all of the paraphrases truly substitutable or do they sometimes exhibit other types of relationships (like directional entailment)? When the paraphrases share the same meaning, are there stylistic reasons why we should choose one versus another (e.g., is one paraphrase a less formal version of another)?

ranked paraphrases of *berries*				
	PPDB 1.0		PPDB 2.0	
1.	embayments	1.	strawberries	$\sqsubset$
2.	strawberries	2.	raspberries	$\sqsubset$
3.	racks	3.	blueberries	$\sqsubset$
4.	grains	4.	blackberries	$\sqsubset$
5.	raspberries	5.	fruits	$\sqsupset$
6.	blueberries	6.	fruit	$\sqsupset$
7.	fruits	7.	beans	#
8.	fruit	8.	grains	$\sim$
9.	blackberries	9.	seeds	#
10.	beans	10.	kernels	#

Figure 1: PPDB 2.0 includes an improved scoring model for ranking paraphrases. Shown are the top 10 ranked paraphrases for the word *berries* according to PPDB 1.0 (left) and PPDB 2.0 (right). PPDB 2.0 also contains an entailment relation for every pair. These relations capture asymmetries in the paraphrases, such as the fact that *strawberries* entails ($\sqsubset$) *berries*, while *fruits* is entailed by ($\sqsupset$) *berries*.

In this paper we describe several improvements to PPDB that address these questions. We release PPDB version 2.0, incorporating the following improvements:

- A completely re-ranked set of paraphrases that uses a regression model to fit the paraphrase scores to human judgments of paraphrase quality. Figure 1 shows the re-ranked paraphrases for the word *berries*.
- Each paraphrase pair is automatically labeled with an explicit entailment relationship. Instead of assuming all paraphrases are perfectly equivalent, we label some as one directional entailments (or other entailment types).
- Each paraphrase rule now has new features that indicate when its application is expected to result in a change in style.
- Each paraphrase entry in the database now has an associated word embedding learned using Multiview Latent Semantic Analysis.

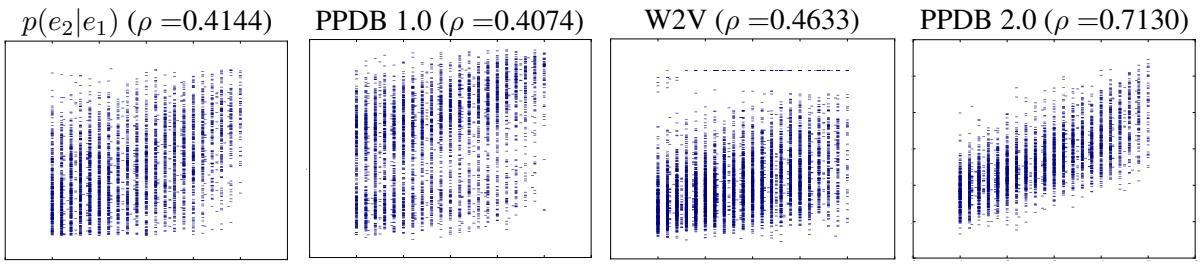

Figure 2: Scatterplots of automatic paraphrase scores (vertical axis) versus human scores (horizontal axis) for four ways of automatically ranking the paraphrases: $p(e_2|e_1)$ (far left), PPDB 1.0's heuristic ranking method (middle left), word2vec similarity (middle right), and our supervised model for PPDB 2.0 (far right). Our rankings achieve the highest correlation with human judgements with a Spearman's ρ of 0.71.

Upon publication of this paper, we will release PPDB 2.0 along with a set of 26K phrase pairs annotated with human similarity judgments.

2 Improved rankings of paraphrases

The notion of ranking paraphrases goes back to the original method that PPDB is based on. Bannard and Callison-Burch (2005) introduced the bilingual pivoting method, which extracts *incarcerated* as a potential paraphrase of *put in prison* since they are both aligned to *festgenommen* in different sentence pairs in an English-German bitext. Since *incarcerated* aligns to many foreign words (in many languages) the list of potential paraphrases is long. Paraphrases vary in quality since the alignments are automatically produced and noisy. In order to rank the paraphrases, Bannard and Callison-Burch (2005) define a paraphrase probability in terms of the translation model probabilities $p(f|e)$ and $p(e|f)$:

$$p(e_2|e_1) \approx \sum_f p(e_2|f)p(f|e_1). \quad (1)$$

Heuristic scoring in PPDB 1.0 Instead of ranking the paraphrases with a single score, Ganitkevitch et al. (2013) expanded the set of scores in PPDB. Each paraphrase rule in PPDB consists of four components: a phrase (e_1), a paraphrase (e_2), a syntactic category (LHS[1]), and a feature vector. This feature vector contains 33 scores of paraphrase quality, which are described in full in the supplementary material to this paper. The rules in PPDB 1.0 were scored using an ad-hoc weighting of seven of these features, given by the following equation:

$$\begin{array}{rrcl} & 1.0 & \times & -log\ p(e_1|e_2) \\ + & 1.0 & \times & -log\ p(e_2|e_1) \\ + & 1.0 & \times & -log\ p(e_1|e_2, LHS) \\ + & 1.0 & \times & -log\ p(e_2|e_1, LHS) \\ + & 0.3 & \times & -log\ p(LHS|e_1) \\ + & 0.3 & \times & -log\ p(LHS|e_2) \\ + & 100 & \times & RarityPenalty \end{array}$$

where $-log\ p(e_2|e_1)$ is the paraphrase probability computed according to Equation 1 and $RarityPenalty$ is a real-valued feature that indicates how frequently the paraphrase was observed in the training data.

This heuristic linear combination of scores was used to divide PPDB into six increasingly large sizes– S, M, L, XL, XXL, and XXXL. PPDB-XXXL contains all of the paraphrase rules and has the highest recall, but the lowest average precision. The smaller sizes contain better average scores but offer lower coverage. Ganitkevitch et al. (2013) performed a small-scale analysis of how their heuristic score correlated with human judgments by collecting <2,000 judgments for PPDB paraphrases of verbs that occurred in Propbank.

Supervised scoring model For this paper, we rank the paraphrases using a supervised scoring model. To train the model, we collected human judgements for 26,455 paraphrase pairs sampled from PPDB. Each paraphrase pair was judged by 5 people who each assigned a score on a 5-point Likert scale, as described in Callison-Burch (2008). These 5 scores were averaged.

We used these human judgments to fit a regression to the 33 features available in the PPDB 1.0 feature vector, plus an additional 176 new features that we developed. Our features included the cosine similarity of the word embeddings that we generated for each PPDB phrase (described in Section 3.3), as well as lexical overlap features, features derived from WordNet, and distributional

[1]The name LHS is due to the fact that the syntactic category comes from the lefthand side of the synchronous CFG rule used to produce the paraphrase.

similarity features. We weighted the contribution of these features using ridge regression with its regularization parameter tuned using cross validation on the training data.

See the supplemental materials for a complete description of the features used in our model and our data collection methodology including inter-annotator agreement.

2.1 Evaluating the rankings

We evaluate the new rankings in two ways:

- We calculate the correlation of the different ways of automatically ranking the paraphrases against the 26k human judgments that we collected.
- We compute the goodness (in terms of mean reciprocal rank and averaged precision) of the ranked paraphrase lists for 100 phrases drawn randomly from Wikipedia.

Correlation Figure 2 plots the different automatic paraphrase scores against the 5-point human judgments for four different ways of ranking the paraphrases: 1) the original paraphrase probability defined by Bannard and Callison-Burch (2005), 2) the heuristic ranking that Ganitkevitch et al. (2013) defined for PPDB 1.0, 3) the cosine similarity of word2vec[2] embeddings[3], and 4) the new score predicted by our discriminative model. The paraphrase probability has a Spearman correlation of 0.41. The heuristic PPDB 1.0 ranking has a similar correlation of $\rho = 0.41$. The word2vec similarity improves correlation slightly to 0.46. To test our supervised method, we use cross validation: in each fold, we hold out 200 phrases along with all of their associated paraphrases for testing. Our rankings for PPDB 2.0 dramatically improve correlation with human judgments to $\rho = 0.71$.

Goodness of the top-ranked paraphrases In addition to calculating the correlation over the sample of paraphrases (where the human judgments were taken evenly over the range of $p(e_2|e_1)$ values), we also evaluated the full list of paraphrases as it is likely to be used by researchers who use PPDB. We took a sample of 100 unique phrase types from Wikipedia (constraining to types which appear in PPDB), and collected human judgments for their full list of paraphrases.

[2] https://code.google.com/p/word2vec/

[3] For phrases, we use the vector of the rarest word as an approximation of the vector for the phrase.

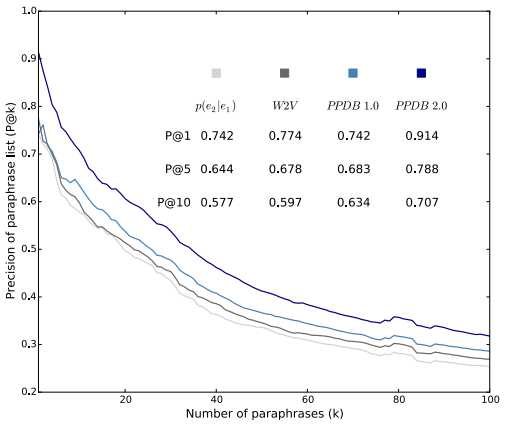

Figure 3: Averaged precision of paraphrases lists for 100 phrases randomly drawn from Wikipedia. Curves show precision @ k for varying values of k, up to 100. Here, "good" paraphrases are defined as having received an average human rating ≥ 3.

		MRR	AP
human rating $\geq$3	Random	0.56	0.46
(16% of judgments)	$p(e_2\|e_1)$	0.84	0.61
	W2V	0.85	0.64
	PPDB 1.0	0.86	0.64
	PPDB 2.0	**0.95**	**0.72**
human rating $\geq$4	Random	0.34	0.27
(4% of judgments)	$p(e_2\|e_1)$	0.69	0.46
	W2V	0.69	0.49
	PPDB 1.0	0.70	0.50
	PPDB 2.0	**0.80**	**0.59**
human rating $\geq$4.5	Random	0.25	0.20
(1% of judgments)	$p(e_2\|e_1)$	0.46	0.37
	W2V	0.46	0.36
	PPDB 1.0	0.53	0.42
	PPDB 2.0	**0.61**	**0.49**

Table 1: Quality of rankings using for the improved PPDB 2.0 score versus the current heuristic score. Both metrics (AP and MRR) range from 0 to 1 and higher is better. $\geq$t means that the statistics are computed by considering a paraphrase to be "good" if its human judgments averaged $\geq$t.

We compare the ranking produced by the proposed PPDB 2.0 model against the heuristic PPDB 1.0 ranking in terms of each one's ability to put good paraphrases at the top of its list. Figure 3 shows precision curves for the ranked paraphrases in PPDB 1.0 compared to PPDB 2.0. PPDB 2.0 achieves consistently higher precision, improving P@1 by 17 points and P@5 by 9 points.

We also analyzed the different rankings when we varied the criterion that we used for what constitutes a good paraphrase. Table 1 shows how the averaged precision (AP) and the mean reciprocal rank (MRR) change as we vary the human score for good paraphrases from $\geq$3 to $\geq$4.5. Depending on the threshold, our PPDB 2.0 ranking

achieves a 9-12 point improvement in MRR over the PPDB 1.0 rankings. Similarly, it improves AP by 7-9 points.

3 Other Additions

In addition to dramatically improving the rankings of the paraphrases (novel to this publication), our PPDB 2.0 release adds several automatic annotations created in other research. Every paraphrase pair now has an entailment relation from Pavlick et al. (2015), style classifications from Pavlick and Nenkova (2015), and associated vector embedding from Rastogi et al. (2015). These are described briefly below.

3.1 Entailment relations

Although we typically think of paraphrases as equivalent or as bidirectionally entailing, a substantial fraction of the phrase pairs in PPDB exhibit different entailment relations. Figure 1 gives an example of how these relations capture the range or entailment present in the paraphrases of *berries*. We automatically annotate each paraphrase rule in PPDB with an explicit entailment relation based on *natural logic* (MacCartney, 2009). These relations include forward entailment/hyponym ($\sqsubset$), reverse entailment/hypernym ($\sqsupset$), non-entailing topical relatedness ($\sim$), unrelatedness ($\#$), and even exclusion/contradiction ($\neg$). For a complete evaluation of the entailment classifications, and the prevalence of each type in PPDB, see Pavlick et al. (2015).

3.2 Style scores

Some of the variation within paraphrase sets can be attributed to stylistic variations of language. We automatically induce style information on each rule in PPDB for two dimensions– complexity and formality. Table 2 shows some paraphrases of *the end*, sorted from most complex to most simple using these scores. These classifications could be useful for natural language generation tasks like text simplification (Xu et al., 2015). A complete evaluation of these scores is given in Pavlick and Nenkova (2015).

3.3 Multiview LSA vector embeddings

Recently there has been tremendous interest in representing words via vector embeddings (Dhillon et al., 2011; Mikolov et al., 2013; Pennington et al., 2014). Such representations can be used to measure word and phrase similarity, possibly to improve paraphrasing. Multiview Latent Semantic Analysis (MVLSA) is a state-of-the-art method for modeling word similarities. MVLSA can incorporate an arbitrary number of data views, such as monolingual signals, bilingual signals, and even signals from other embeddings. PPDB 2.0 contains new similarity features based on MVLSA embeddings for all phrases. A complete discussion is given in Rastogi et al. (2015).

1. the finalization	6. the latter part	11. the final analysis
2. the expiration	7. termination	12. the last
3. the demise	8. goal	13. the finish
4. the completion	9. the close	14. the final part
5. the closing	10. late	15. the last part

Table 2: Some paraphrases of *the end*, ranked from most complex to most simple according to the style scores included in PPDB 2.0.

4 Related Work

The most closely related work to our supervised re-ranking of PPDB is work by Zhao et al. (2008) and Malakasiotis and Androutsopoulos (2011). Zhao et al. (2008) improved Bannard and Callison-Burch (2005)'s paraphrase probability by converting it into log-linear model inspired by machine translation, allowing them to incorporate a variety of features. Malakasiotis and Androutsopoulos (2011) developed a similar model trained on human judgements. Both efforts apply their model to natural language generation by paraphrasing full sentences. We apply our model to the sub-sentential paraphrases directly, in order to improve the quality of the Paraphrase Database.

Also related is work by Chan et al. (2011) which reranked bilingually-extracted paraphrases using monolingual distributional similarities, but did not use a supervised model. Work that is relevant to our classification of semantic entailment types to each paraphrase, includes learning directionality of inference rules (Bhagat et al., 2007; Berant et al., 2011) and learning hypernyms rather than paraphrases (Snow et al., 2004). Our style annotations are related to Xu et al. (2012)'s efforts at learning stylistic paraphrases. Our word embeddings additions to the paraphrase database are related to many current projects on that topic, including projects that attempt to customize embeddings to lexical resources (Faruqui et al., 2015). However, the Rastogi et al. (2015) embeddings included here were shown to be state-of-the art in

predicting human judgements.

5 Conclusion

We release PPDB 2.0 (`http://paraphrase.org/#/download`). The resource includes dramatically improved paraphrase rankings, explicit entailment relations, style information, and state-of-the-art distributional similarity measures for each paraphrase rule. The 2.0 release contains 100m+ paraphrases, and 26k manually rated phrase pairs, which will facilitate further research in modeling semantic similarity.

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 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.

We would like to thank the anonymous reviewers for their thoughtful comments.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL*, pages 597–604.

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

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *ACL*.

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.

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. In *SemEval*.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *EMNLP*, pages 196–205.

Tsz Ping Chan, Chris Callison-Burch, and Benjamin Van Durme. 2011. Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *GEMS*, pages 33–42.

Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. 2011. Multi-view learning of word embeddings via CCA. In *NIPS*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL*.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *EMNLP*, pages 1168–1179.

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

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. Umbc ebiquity-core: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 44–52.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *EMNLP*, pages 891–896.

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

Prodromos Malakasiotis and Ion Androutsopoulos. 2011. A generate and rank approach to sentence paraphrasing. In *EMNLP*, pages 96–106.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *NAACL*.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA: Representation learning via generalized CCA. In *NAACL*.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*.

Md. Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014b. Dls@cu: Sentence similarity from word alignment. In *SemEval*.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *TACL*.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Semi-markov phrase-based monolingual alignment. In *EMNLP*, pages 590–600.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *ACL*, volume 2, pages 545–550.

Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. 2008. Combining multiple resources to improve SMT-based paraphrasing model. In *ACL*.