Comparing Constraints for Taxonomic Organization: Supplementary Materials

Anne Cocos[*], **Marianna Apidianaki**[*†], and **Chris Callison-Burch**[*]

[*] Department of Computer and Information Science, University of Pennsylvania
[†] LIMSI, CNRS, Université Paris-Saclay, 91403 Orsay
{acocos,marapi,ccb}@seas.upenn.edu

1 Introduction

Here we provide further details on our PPDB Local Taxonomies dataset and relation prediction.

1.1 PPDB Local Taxonomies Dataset

In Table 1, we list the 50 target nouns and the size of each of their local PPDB taxonomies.

Target (# terms)		
Validation Set		
chip (14)	dealer (48)	mark (32)
note (60)	reputation (37)	
Test Set		
access (46)	accounting (34)	address (35)
air (46)	body (46)	camp (19)
campaign (39)	cell (21)	challenge (53)
class (36)	commission (21)	community (63)
display (50)	edge (49)	entry (84)
failure (126)	field (76)	flight (54)
foundation (41)	function (41)	gap (88)
gas (22)	guarantee (34)	house (52)
idea (97)	innovation (24)	legislation (33)
margin (33)	market (37)	mind (90)
moment (34)	movement (57)	office (52)
officer (58)	origin (29)	park (13)
promotion (74)	rally (37)	road (58)
screen (38)	shape (25)	speed (39)
television (16)	threat (29)	tour (47)

Table 1: Target nouns and number of terms in the corresponding PPDB local taxonomy.

1.2 Relation Prediction

Our dataset for training and evaluating the HypeNET hypernym classifier consists of two portions: a benchmark portion, and a PPDB portion. The details of how we construct the dataset are as follows.

The benchmark portion of our dataset contains pairs from four existing datasets: BLESS (Baroni and Lenci, 2011), ROOT09 (Santus et al., 2016), EVALution (Santus et al., 2015), and K&H+N (Necsulescu et al., 2015) (an extension of the Kozareva and Hovy (2010) dataset).[1] Each of these is a multi-class relation prediction dataset; we binarize the data by labeling noun pairs with a hypernym-like relation as positive instances, and all others as negative. We add the training and validation splits for each dataset (78,144 noun pairs in all) to our training set, and the test splits (26,051 noun pairs total) to our test set. In total, 10.3% of the benchmark training instances and 10.1% of test instances are hypernyms.

To the benchmark datasets we add a set of noun pairs extracted from PPDB, where the ground truth relation labels are derived from WordNet 3.0 (Miller, 1995). The noun pairs can consist of single- or multi-word phrases. This portion consists of 50,402 related noun pairs that appear in PPDB and hold a direct synonym, hypernym, hyponym, or meronym relation in WordNet, plus an additional 46,908 noun pairs that are unrelated in WordNet (but share at least one PPDB paraphrase in common), for a full (unfiltered) dataset size of 97,310 noun pairs. The unrelated pairs are carefully selected to reflect the type of unrelated pairs we expect to see when creating local PPDB taxonomies – namely, word pairs that have at least one PPDB paraphrase in common, but which are (a) not linked in WordNet, and (b) not directly linked in PPDB. From these 97k pairs, we randomly extract 1,000 unrelated and 2,000 related noun pairs to add to our held-out test set, and the remaining 77,822 pairs that share no words in

[1] These datasets have been compiled by Shwartz and Dagan (2016b) and can be found here: https://github.com/vered1986/LexNET

Dataset	# Test Instances	% Hypernym	P-hyper	R-hyper	F1-hyper	P-avg	R-avg	F1-avg
PPDB	3,000	20.1	.443	.476	.459	.781	.774	.777
BLESS	6,637	5.27	.723	.903	.803	.980	.977	.978
EVALution	1,846	24.54	.491	.764	.598	.804	.748	.763
K&H+N	14,377	7.25	.888	.950	.918	.988	.988	.988
ROOT09	3,191	26.25	.580	.752	.655	.825	.800	.808

Table 2: Evaluation of the HypeNET hypernym classifier on the PPDB test set and four benchmark test sets. Label-specific scores report precision, recall, and F1-Score for just hypernyms; the weighted average combines these metrics for hypernyms and non-hypernyms, with each weighted by the number of that type in the evaluation set.

Dataset	# Test Instances	% Synonym	P-syn	R-syn	F1-syn	P-avg	R-avg	F1-avg
PPDB	3,000	24.65	.438	.237	.307	.698	.737	.707
EVALution	1,846	15.1	.325	.278	.300	.792	.804	.797

Table 3: Evaluation of the synonym classifier on the PPDB test set and EVALution benchmark set (the only benchmark containing synonyms). As in Table 2, we give scores in terms of synonym-specific and weighted average precision, recall, and F1-Score.

common with the test pairs for training and validation. Again we binarize the dataset, using the 10.9% of training and 20.1% of test pairs having a hypernym relation as positive instances. We ensure lexical separation from our taxonomy induction dataset; no terms in the classifier training set appear in any of the local taxonomies.

After we combine the benchmark and PPDB datasets, we do further trimming to maintain a 1:4 positive:negative class ratio in the training set as was done by Shwartz et al. (2016). The combined benchmark+PPDB training set of 149,334 unique pairs has only 12.7% hypernyms, so we take all positive hypernym pairs from the combined set and randomly choose enough negative pairs from the remainder to maintain the 1:4 ratio. The final training set has 76,152 noun pairs.

We train and validate HypeNET using our 76K-pair test set, and use the trained model to predict hypernym relations for every pair of terms that appear in the same PPDB local taxonomy. To assess the strength of the hypernym prediction model, we also predict relations for the noun pairs in our benchmark and PPDB test sets. The results, given in Table 2, are not directly comparable to those reported in Shwartz and Dagan (2016a) because we have binarized the labels, but fall roughly within the same range for each dataset.

We similarly evaluate our synonym prediction method, which predicts any two terms having a PARAGRAM vector cosine similarity of at least 0.76 to be synonyms, on the PPDB and EVALution datasets. Results are in Table 3.

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Edinburgh, Scotland, pages 1–10.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on Empirical Methods in Natural Language Processing (EMNLP)*. Cambridge, Massachusetts, USA, pages 1110–1118.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Silvia Necsulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. 2015. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM)*. Denver, Colorado, USA, pages 182–192.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine features in a random forest to learn taxonomical semantic relations. Portoroz, Slovenia, pages 4557–4564.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*. Beijing, China, pages 64–69.

Vered Shwartz and Ido Dagan. 2016a. Cogalex-v shared task: LexNET - Integrated path-based and distributional method for the identification of semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*. Osaka, Japan, pages 80–85.

Vered Shwartz and Ido Dagan. 2016b. Path-based vs. distributional information in recognizing lexical semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*. Osaka, Japan, pages 24–29.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany, pages 2389–2398.