Classification of Entailment Relations in PPDB

1 Overview

This document outlines our protocol for labeling noun pairs according to the entailment relations proposed by Bill MacCartney in his 2009 thesis on Natural Language Inference. Our purpose of doing this is to build a labelled data set with which to train a classifier for differentiating between these relations. The classifier can be used to assign probabilities of each relation to the paraphrase rules in PPDB, making PPDB a more informative resource for downstream tasks such as recognizing textual entailment (RTE).

2 Entailment Relations

MacCartney’s thesis proposes a method for partitioning all of the possible relationships that can exist between two phrases into 16 distinct relationships. Of these, he defines the 7 non-degenerate relations as the “basic entailment relationships.” MacCartney’s table explaining these 7 relations is reproduced in figure 1.

![Table of MacCartney's relations](image)

In summary, the relations are:

- **equivalence (=)**: if X is true then Y is true and if Y is true then X is true.

- **forward entailment (<)**: if X is true then Y is true but if Y is true then X may or may not be true.

- **reverse entailment (>)**: if Y is true then X is true but if X is true then Y may or may not be true.

- **negation (\(\land\))**: if X is true then Y is not true and if Y is not true then X is true, and either X or Y must be true.

- **alternation (\(|\))**: if X is true then Y is not true but if Y is not true then X may or may not be true.

- **independence (\(\#\))**: if X is true then Y may or may not be true and if Y is true then X may or may not be true, and either X or Y must be true.
may not be true and if Y is true then X may or may not be true.

As MacCartney acknowledges, the utility of the cover relation for RTE is “not immediately obvious,” and we do not use this relation in our work.

3 Defining relations within Wordnet

Below, we describe our rules for mapping a pair of nodes in the Wordnet noun hierarchy onto one of the six basic entailment relations. Our mappings are shown pictorially in figure 2.

We use Python NLTK’s Wordnet interface with Wordnet version 3.1 to implement the described algorithms.

In the following definitions, lowercase letters represent strings and capital letters represent Wordnet synsets.

**Synonym**  
X and Y are considered synonyms if any sense of X shares a synset with any sense of Y. That is, if

\[ \text{synsets}(X) \cap \text{synsets}(Y) \neq \emptyset \]

**Hypernym**  
X is considered a hypernym of Y if some synset of X is the root of a subtree containing some synset of Y. I.e.,

\[ \text{synsets}(Y) \cap \bigcup_{X \in \text{synsets}(X)} \text{children}(X) \neq \emptyset \]

where \( \text{children}(X) \) is the transitive closure of the children of X (i.e. all the nodes in the subtree rooted at X).

**Hyponym**  
X is considered a hyponym of Y if Y is a hypernym of X, using the definition of hypernym given above. Figure 3 gives a visual representation of how this process establishes that “drop” is a hyponym of “amount.”

**Antonym**  
Wordnet defines the antonym relationship over lemmas, rather than synsets. We therefore define the set of antonyms of X as

\[ \text{antonyms}(X) = \bigcup_{X \in \text{synsets}(X)} \bigcup_{w_x \in X} \text{WNantonyms}(w_x) \]

where \( w_x \) is a lemma and \( \text{WNantonyms}(w_x) \) is the set of lemmas Wordnet defines as antonyms for \( w_x \).

We then call X and Y antonyms if \( X \in \text{antonyms}(Y) \) or \( Y \in \text{antonyms}(X) \).

**Alternation**  
We say that X and Y are in alternation if X and Y are in disjoint subtrees rooted at a shared hypernym. I.e., X and Y are in alternation if both

1. None of the conditions for synonym, hypernym, hyponym, or antonym are met, and
2. \( \text{hypernyms}(X) \cap \text{hypernyms}(Y) \neq \emptyset \).

Wordnet is structured so that the entire noun hierarchy is rooted at “entity.” The hierarchy is fairly fat, with most nodes having a depth between 5 and 10 (figure 5). To avoid having too many false positives for the alternation class, we do not count two words as alternation if their deepest shared hypernym is in the first three levels of the hierarchy (i.e. has a depth less than three). This amounts to removing 26 synsets as potential shared hypernyms for alternation, which are shown in table 5. These nodes largely represent abstract categories such as “physical entity” and “causal agent.” Removing these nodes prevents true independent pairs from being falsely labeled as alternation. We elaborate on the more potential false positives for the alternation class in section 7.

**Other Relatedness**  
In addition to MacCartney’s relations, we define a sixth relation category, which
serves as a catch-all for terms which are flagged as related by WordNet but whose relation is not built into the hierarchical structure. WN refers to these relations as either semantic, meaning they exist as pointers between synsets, or as lexical, meaning they exist as pointers between lemmas.

While these noun pairs do not meet the criteria of the basic entailment relations defined, we feel they carry more information about entailment than do truly independent terms. Currently, we combine all of these forms of relatedness into an “other” category, and we leave the exact entailment implications of this category to be learned in downstream applications.

**Semantic relations** We use two of WN’s semantic noun-noun relations, holonymy and meronymy. WN breaks each of these into three subclasses: part, member, and substance. For example, arm/body is classified as part holonym, musician/musical group is classified as member holonym, water/ice is classified as substance holonym. We combine these three fine-grained classes and use only the coarser holonym and meronym labels. We define \( x \) and \( y \) as holonyms analogously to our definition of hyponymy:

\[
\text{synsets}(y) \cap \bigcup_{X \in \text{synsets}(x)} \text{WNholonyms}(X) \neq \emptyset
\]

where \( \text{WNholonym}(X) \) is the transitive closure of the holonyms of \( X \) according to WN. Meronymy is defined identically.

**Lexical relations** We use two of WN’s lexical relations: derivational relatedness and pertainym. Pertainym pointers exist between adjectives and nouns, not between nouns and other nouns. Because PPDB contains unreliable part of speech labels, we assign the pertainym labels to describe the relation between a pair of nouns \( x/y \) if \( x \) is lexically identical to an adjective reachable from \( y \) through a pertainym pointers, or vis-versa. We define pertained of \( x \) analogously to antonyms:

\[
\text{pertain}(x) = \bigcup_{X \in \text{synsets}(x)} \bigcup_{w_x \in X} \text{WNpertain}(w_x).
\]

We define the lemmas of \( y \) as:

\[
\text{lemmas}(y) = \bigcup_{Y \in \text{synsets}(y)} \bigcup_{w_y \in Y} w_y.
\]

We then call \( y \) a pertainym of \( x \) if \( \text{lemmas}(y) \cap \text{pertain}(x) \neq \emptyset \).

Derivational relatedness is defined identically, with a slight modification to enforce transitivity. I.e. \( x \) is derivationally related to \( y \) if \( \text{lemmas}(y) \cap \text{deriv}(x) \neq \emptyset \) or \( \text{deriv}(x) \cap \text{derive}(y) \neq \emptyset \). This is to deal with a small number of cases in which two nodes share a derivational stem but lack an explicit pointer between them, e.g. WN relates vaccine/vaccinate and vaccination/vaccinate, but not vaccine/vaccination.

**Semantic relations treated as lexical relations** We also use WN’s attribute pointer, which is given as a semantic relation (between synsets) but exists as a pointer from an adjective synset to a noun synset. We want to treat this relation as we did with pertainym, that is we want to assign the attribute label to noun pairs if \( x \) is lexically identical to an adjective reachable from \( y \) through an attribute pointer, or vis-versa. We therefore treat attribute as a lexical relation by taking the union over the lemmas of synsets, rather than over synsets:

\[
\text{attr}(x) = \bigcup_{X \in \text{synsets}(x)} \bigcup_{A \in \text{WNattr}(X)} \bigcup_{w_x \in A} w_x
\]

where \( \text{WNattr}(x) \) is the set of synsets for which WN considers \( X \) to be an attribute. As before, we then call \( y \) an attribute of \( x \) if \( \text{lemmas}(y) \cap \text{attr}(x) \neq \emptyset \).

**Independent** Independence is simply defined as failing to meet any of the criteria for the above 5 basic entailment relations or for the “other” category.

### 4 Preprocessing steps

We remove noun pairs in which \( x \) and \( y \) are inflected forms of the same lemma. In querying for WordNet synsets, however, we leave each noun in its inflected form. This is to address the fact that NLTK lemmatizes automatically when no synset exists for the inflected form (e.g. it will map “dogs” to the synset for “dog”) but requires inflections when there
Table 1: Some example noun pairs for each label defined according to Wordnet.

<table>
<thead>
<tr>
<th>independent</th>
<th>synonym</th>
<th>antonym</th>
<th>hypernym</th>
<th>hyponym</th>
<th>independent</th>
<th>other_related</th>
<th>synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>actress/suspect</td>
<td>antonym/simile</td>
<td>amount/droplet</td>
<td>aryan/being</td>
<td>appeal/</td>
<td>antuala/bosphoros</td>
<td>belongings/property</td>
<td></td>
</tr>
<tr>
<td>contamination/desication</td>
<td>assembly/disassembly</td>
<td>box/casket</td>
<td>battle/ground/region</td>
<td>armenian/consonant</td>
<td>blue/</td>
<td>centers/plaza</td>
<td></td>
</tr>
<tr>
<td>cup/dons</td>
<td>bemignity/malignancy</td>
<td>care/medication</td>
<td>blatteria/order</td>
<td>cops/machination</td>
<td>champion/mount</td>
<td>dispatch/expedition</td>
<td></td>
</tr>
<tr>
<td>disguise/monsignor</td>
<td>broadening/narrowing</td>
<td>content/formalism</td>
<td>botswana/countries</td>
<td>farewell/string</td>
<td>dissent/</td>
<td>environment/surroundings</td>
<td></td>
</tr>
<tr>
<td>dory/race</td>
<td>end/onset</td>
<td>country/kabul</td>
<td>carol/vocals</td>
<td>freshness/obligation</td>
<td>fjiyama/japan</td>
<td>lasagna/asagne</td>
<td></td>
</tr>
<tr>
<td>enterprise/usasa</td>
<td>evil/goodness</td>
<td>country/uganda</td>
<td>connecticut/district</td>
<td>general/inspection</td>
<td>halotida/halotis</td>
<td>level/leveler</td>
<td></td>
</tr>
<tr>
<td>governance/presentation</td>
<td>foes/friend</td>
<td>district/omaha</td>
<td>family/parentage</td>
<td>hitchcock/vocals</td>
<td>mounain/pack</td>
<td>pa/pap</td>
<td></td>
</tr>
<tr>
<td>magazine/tricycle</td>
<td>hardness/softness</td>
<td>genus/populus</td>
<td>fashion/practice</td>
<td>ko/state</td>
<td>quivering/trembling</td>
<td>sentence/synonym</td>
<td></td>
</tr>
<tr>
<td>phenomenon/booth</td>
<td>hyperatremia/hyperatremia</td>
<td>ground/she/now/land</td>
<td>fisher/workers</td>
<td>south/westen/special</td>
<td>wave/weiss</td>
<td>yankess/yanks</td>
<td></td>
</tr>
<tr>
<td>region/science</td>
<td>imprecision/precision</td>
<td>object/wedge</td>
<td>male/someone</td>
<td>wave/waves</td>
<td>stoppage/stopping</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Distribution of labels over 20,624,173 noun pairs found in Wordnet.

is a differentiation in the synsets (e.g. querying for “marine” will return the synset for “a member of the United States Marine Corps” but only querying for the pluralized “marines” will return the synset for “a division of the United States Navy.”)

We check the criteria for each label in sequence, to enforce mutual exclusivity of labels. The order in which the criteria are checked is 1) antonym, 2) synonym, 3) hypernym, 4) hyponym, 5) other relatedness, 6) alternation, 7) independence. This means that if a noun pair fits the criteria for multiple relations (due to multiple senses, for example), the labels will be prioritized in this order. For example, the noun pair king/queen fits the criteria for antonym as well as for synonym (under the synset “a competitor who holds a preeminent position”), but it is assigned only to the antonym relation.

5 Results

We ran the described labeling method on 20,624,173 noun pairs. The nouns in each pair occurred together in a sentence in the Wackypedia corpus and were joined by a path of no more than 5 nodes in the parse tree. Table 1 shows some sample noun pairs and table 2 gives the number of noun pairs matching each relation. Table 4 shows a sample of the labels assigned to noun pairs from occurring both in Wackypedia and in PPDB.

6 Weaknesses and limitations

Deviations from MacCartney’s definitions The structure of the Wordnet hierarchy and our definitions result in some differences between our classes and those proposed by MacCarthy. Whereas MacCarthy’s alternation is a strictly contradictory relationship (if X is true, than Y cannot be true), our alternation class does not have this property.

Wordnet’s inconsistent use of the parent/child pointers prevents us from consistently identifying when a shared parent node indicates true alternation verses when it indicates a hypernymy relationship. Figure 4 shows an example of how the relationship between a node and its immediate children may exemplify either relationship. Because we cannot systematically minimize these errors (we elaborate in section 7), we do nothing to prevent this form of false positive. We recognize that, compared to the strong exclusivity implication of MacCartney’s alternation (table 3), our alternation will be weaker in RTE tasks.

Additionally, MacCartney’s definition of antonym requires that antonym pairs have the “exhaustiveness” property; that is, it requires that if x and y are antonyms, then any noun must either be x or be y. This is not the definition of antonym used by Wordnet. For example, Wordnet provides the antonym pair man/woman, even though it is possible for some noun to be neither a man nor a woman. Because of this, many of our pairs labelled antonym more accu-
Table 3: Our definition of alternation provides substantially weaker signal for entailment recognition than does MacCartney’s alternation.

<table>
<thead>
<tr>
<th>Entailment</th>
<th>MacCartney</th>
<th>Wordnet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Containment</td>
<td>equivalence</td>
<td>synonym</td>
</tr>
<tr>
<td></td>
<td>forward entailment</td>
<td>hyponym</td>
</tr>
<tr>
<td></td>
<td>reverse entailment</td>
<td>hypernym</td>
</tr>
<tr>
<td>Contradiction</td>
<td>negation</td>
<td>antonym</td>
</tr>
<tr>
<td>Compatibility</td>
<td>independent</td>
<td>alternation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>independent</td>
</tr>
</tbody>
</table>

Table 4: Inconsistent levels of granularity in the Wordnet hierarchy can cause containment relations (hyponym/hypernym) to be falsely labeled as alternation. On the left is true alternation. On the right, the child nodes would be most appropriately labeled as hypernym.

Figure 4: Figure showing inconsistent levels of granularity in the Wordnet hierarchy can cause containment relations (hyponym/hypernym) to be falsely labeled as alternation. On the left is true alternation. On the right, the child nodes would be most appropriately labeled as hypernym.

Other limitations We try to work only with nouns. We therefore draw all of our labels from WN’s noun hierarchy and apply our labeling only to phrases in PPDB which are tagged as nouns. The tagging in PPDB is not perfect, and we receive some supposed noun/noun pairs which are more likely a different part of speech (e.g. basic/fundamental). Classifying these pairs under the assumption that they are nouns causes noisy labeling (e.g. basic/fundamental is labeled as independent).

Wordnet has peculiarities which sometimes cause incorrect labels. One issue is that of word sense. Despite Wordnet’s effort to cover a large number of senses for each noun, it is not always complete. For example, Wordnet causes the pairs constraint/limitation and downturn/slowdown to be labeled as alternation, even though they are most often used synonymously. Another issue is that of over-formalization, which causes a disconnect between the information in Wordnet and the way terms are used colloquially. For example, Wordnet defines objectivity as an “attribute” and impartiality as a “psychological feature,” thus causing the pair objectivity/impartiality to be labeled as independent, despite the fact that these two terms are used interchangeably. Similarly, Wordnet defines “dairy” as a hyponym of “farm” and “dairy product” as a hyponym of “food,” causing the pair dairy/milk to be labeled as independent.

7 Handling alternations

As discussed above, the alternation class presents some particular difficulties when defined using only the information available in Wordnet. In this section, we explore in detail the shortcomings of the Wordnet approach and motivate our decision to collect manual annotations for our data.

7.1 Definitions

We refer to the depth of a node as the length of the path from that node to the root. Some nodes have multiple inheritance or are reachable from the root node via more than one path. The NLTK Wordnet interface provides access to both the minimum depth (the length of the shortest path from root to node) and the maximum depth (the length of the longest path from root to node). In the following experiments, we mean the minimum depth when we refer to depth. All of the experiments were replicated using the maximum depth, and showed no notable change in results.

When we refer to the distance between two nodes, we are referring to the length of the shortest path between the two nodes.

When we refer to the shared parent of two nouns, we refer to the deepest common hypernym of both nouns. That is, in cases where the nouns share more than one hypernym, we choose the hypernym with the greatest depth. For example, the nouns “ape” and “monkey” share the hypernym “primate,” which has a depth of 11 and also “animal” which has a depth of 6. We choose “primate” as the deepest shared hypernym.

1For example, the synset person.n.01 has a depth of 3 via the path entity→physical entity→causal agent→person, as well as a depth of 6 via the path entity→physical entity→object→whole→living thing→organism→person.
Table 4: Labeling applied to noun pairs in PPDB-large. Only one pair appearing in WN (proliferation/nonproliferation) was assigned the label “antonym.”

7.2 Parameters

We want to optimize the accuracy of classifying a noun pair \( x/y \) as “alternation.” Using only the information within Wordnet, we have three basic parameters to play with: the depth of the \( x \) and \( y \)’s shared parent in the hierarchy, the distance of the shared parent from the either \( x \) or \( y \), and the number of children the parent has.

**Depth of parent**  Wordnet is organized hierarchically, so intuitively, nodes which are nearer to the root refer to more general topics, and nodes deeper in the hierarchy become more specific. Figure 5 shows how the top levels of the hierarchy contain very few nodes, and table 5 shows the kinds of general concepts that exist in these top levels.

It is therefore seems reasonable to assume that, for a noun pair labeled as alternation, the depth of the shared parent may be indicative of the accuracy of the alternation: noun pairs for which the lowest shared parent is very high in the hierarchy (close to the root) may actually be independent (e.g. if both nouns are hyponyms of “thing” but otherwise are unconnected) and noun pairs for which the lowest shared parent is very deep in the hierarchy (close to the leaf nodes) may actually be in a hypernymy relationship (e.g. if both nodes are children of something specific like “happiness”). We test the hypothesis that the depth of the shared parent can be tuned to reduce the number of false positive labels for alternation.

**Distance from parent**  Additionally, we consider the nouns’ distance from their shared parent. While in some alternation pairs, both nouns are immediate decedents of their shared hypernym (e.g. “love” and “fear” are both children of “emotion”), other pairs consist of nouns which are very far from the shared parent (e.g. “step father” and “oil painter” can be classified as alternation through the shared parent “person,” but the nouns are 7 and 4 steps away, respectively). We explore the idea of limiting the maximum distance from the shared parent in order to increase the precision of our alternation classification.

It is worth noting that Wordnet’s “sister terms” are equivalent to our alternation if we restrict the maximum distance from the shared parent to be no more than 1.

**Number of children**  Finally, we consider the number of immediate children of the shared parent. While it is less intuitive than the previous para-
ters how this may affect the likelihood of alternation, it may be that parents with many children are nodes which enumerate subtypes (e.g. the synset for “flower” has over 100 children, which enumerate different types of flowers) and thus are high-precision indicators of alternation. We therefore also investigate the relationship between the number of children of a node and the likelihood of the children being in alternation.

7.3 Labelled data for testing

In order to compare the effect of altering these parameters, we manually labeled 200 noun pairs as alternation/non-alternation. We chose these pairs by randomly sorting the noun pairs that appeared in PPDB (xl) and were labeled as alternation using our original configuration (minimum depth of 3 and no other restrictions). We labelled the first 100 unambiguous examples of alternation and the first 100 unambiguous examples of non-alternation. The full list used is included at the end of this paper (table 8). This data is used to compute precision/recall measures in the following experiments.

7.4 Experiments

We first performed a qualitative analysis. Manually inspecting the data to evaluate the minimum depth parameter, we see that there are examples of both good and bad alternation pairs alternating through parents at every depth. Table 6 illustrates this.

Manual inspection of the maximum distance parameter is similarly inconclusive. We ran our tagging algorithm on a random sample of 30 manually labeled pairs (15 alternation, 15 non-alternation), setting max distance $d$ to each value from 1 to 10, and enforcing the constraint that a noun pair is in alternation if the nouns in the pair share a hypernym and neither noun is more than $d$ hops from the shared hypernym. Increasing $d$ increases equally the number of true positives and the number of false positives. At $d = 7$, all 30 nouns were labeled as alternation. Table 7 shows these results.

We also look quantitatively at the effects of these parameters. Figure 6 shows the precision and recall curves generated by a) fixing the max distance parameter at infinity and varying the min depth between 1 and 10 and b) fixing the min depth at 0 and varying the max distance parameter between 1 and

<table>
<thead>
<tr>
<th>depth</th>
<th>parent</th>
<th>x/y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>whole.n.02</td>
<td>ballplayer/baseball</td>
</tr>
<tr>
<td></td>
<td>state.n.02</td>
<td>bliss/happiness</td>
</tr>
<tr>
<td>4</td>
<td>happening.n.01</td>
<td>avalanche/flood</td>
</tr>
<tr>
<td></td>
<td>disorder.n.03</td>
<td>instability/turmoil</td>
</tr>
<tr>
<td>5</td>
<td>linear_unit.n.01</td>
<td>km/mile</td>
</tr>
<tr>
<td></td>
<td>district.n.01</td>
<td>town/village</td>
</tr>
<tr>
<td>6</td>
<td>meal.n.01</td>
<td>dinner/lunch</td>
</tr>
<tr>
<td></td>
<td>atmospheric_phenomenon.n.01</td>
<td>snow/snowstorm</td>
</tr>
<tr>
<td>7</td>
<td>modulation.n.02</td>
<td>am/pm</td>
</tr>
<tr>
<td></td>
<td>payment.n.01</td>
<td>compensation/wage</td>
</tr>
<tr>
<td>8</td>
<td>garment.n.01</td>
<td>sweater/t-shirt</td>
</tr>
<tr>
<td></td>
<td>man.n.01</td>
<td>boy/guy</td>
</tr>
<tr>
<td>9</td>
<td>movable_feast.n.01</td>
<td>easter/passover</td>
</tr>
<tr>
<td></td>
<td>clergyman.n.01</td>
<td>preacher/priest</td>
</tr>
<tr>
<td>10</td>
<td>bird_of_prey.n.01</td>
<td>eagle/hawk</td>
</tr>
<tr>
<td></td>
<td>placental.n.01</td>
<td>cattle/livestock</td>
</tr>
</tbody>
</table>

Table 6: Pairs of words labeled as alternation, and the words’ shared parent. Parents at every depth can produce both good (top) and bad (bottom) examples of alternation.

10. In the former case (left of figure 6), we see that we are able to achieve acceptable precision when we set a high minimum depth, but only at the expense of unacceptably low recall. In the latter case (right of figure 6), precision stays near 50% (the equivalent of labeling everything as alternation) no matter what value we choose for max distance.

![Figure 6: Precision and recall curves when varying max distance (left) and min depth (right). For max distance curves, min depth was fixed at 0. For min depth curves, max distance was fixed at 20. A precision of 0.5 and a recall of 1.0 can are achieved by labeling all pairs as alternation.](image)

We also look at the effect of using these two parameters (maximum distance from parent and minimum depth of parent) together. Figure 7 shows the F1 scores achieved for different combinations of parameter settings. No combination achieves higher than 0.67, the F1 score achievable by labeling everything as alternation (which results in 50% precision.
Table 7: Noun pairs labeled as alternation (out of a test set of 30 manually labeled pairs) with varying values of max distance. The numbers of true positives and of false positives increase equally as the max distance increases.

Figure 7: F1 scores for varying minimum depths (rows) and maximum distances (columns). Precision and recall were calculated on a set of 200 manually labeled noun pairs (100 positive examples of alternation, 100 negative). No combination achieves acceptable results: 0.67 is the F1 score that would be achieved by labeling every pair as alternation (0.5 precision and 1.0 recall).

Finally, we look at the association between the number of children a parent has and the likelihood that those children are in alternation. Figure 8 shows a histogram of the number of immediate children per parent for each pair in our manually labeled data set. The distributions of number of children is nearly identical for both alternation and non-alternation pairs. We also ran a precision-recall analysis varying both the maximum and the minimum number of children. That is, we tried labeling pairs as alternation if 1) both nouns shared a parent and 2) that parent had at least/at most k children. Although the graphs are not shown, the results were the same as in the graph for maximum distance: for all values of k and for both the “at least” and the “at most” setting, precision remained at 50%, equivalent to random guessing.

Figure 8: Number of immediate children of the deepest shared parent for non-alternation pairs (left) and for alternation pairs (right). The distributions are almost identical, suggesting that the number of children of the shared parent is not a useful way of reducing the number of false positive alternation labels.

8 Next steps

Do to the described limitations, and the goal application to RTE, we will gather manual annotations for a subset of our data in PPDB. By using more reliably labeled data than what we can collect from Wordnet, we should be able to train a classifier to automatically label the remainder of the pairs in PPDB. The
Table 8: Manually annotated noun pairs used in precision/recall calculations.

resulting labels (or distributions over labels) will hopefully offer a valuable contribution to RTE systems, with a scale and level of annotation that is not currently available in existing resources.

9 References