Disentangling Word Association in Distributional Semantics with WordNet

Serina Chang
Department of Computer Science
Stanford University
serinac@stanford.edu

Abstract

The discovery that word distributions encode meaning about words has spurred progress in many areas of NLP. However, we have yet to develop a clear understanding of what kinds of meaning are encoded in distributional models. This oversight may lead to problematic results; for example, prior research has shown that word embeddings encode gender bias and stereotypes of ethnic minorities. In this paper, I begin to disentangle types of meaning by asking, to what extent can word association in distributional models be explained by definitional word association (and what remains beyond definition)? To answer this, I derive parallel graphs from WordNet and word embeddings, and compare the results of semantic field induction over these graphs. I find that definitional association, based on WordNet, can explain a portion (20%-30%) of distributional association, but it is far from telling the whole story. This finding, in addition to fine-grained examples and analyses from each semantic field, brings us closer to pinpointing what is encoded in the embeddings that power our state-of-the-art systems.

1 Introduction

"You shall know a word by the company it keeps" (Firth, 1957). This is the underlying argument of the Distributional Hypothesis, which asserts that a word's meaning can be derived from its distribution in natural language corpora, and words that occur in similar contexts have similar meaning. The Distributional Hypothesis has spurred progress in many areas of natural language processing, as it has created the opportunity to learn word meaning directly from data, instead of relying on manually constructed dictionaries or taxonomies. However, despite the improvements in performance driven by distributional semantics, a significant gap in knowledge remains: what types of word meaning, exactly, do we capture in the company that a word keeps?

Missing the answer to this question can lead to dangerous conclusions. For example, Bolukbasi et al. 2016 show in their seminal work that "even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent." Furthermore, Garg et al. 2018 show that word embeddings encode bias so clearly that we can track how stereotypes towards women and ethnic minorities have evolved in the 20^{th} and 21^{st} centuries simply by training word embeddings on corpora per decade. These findings illustrate that word embeddings encode more than definitional knowledge, and it is paramount that we disentangle what types of knowledge are encoded; otherwise, we run the risk of unintentionally perpetuating biases.

Disentangling word association, however, is an daunting problem. A starting point would be to separate definitional versus non-definitional association: for instance, "woman" and "sister" are associated by definition, while "woman" and "homemaker" are not. A reasonable set-up would be to construct a model of definitional association and compare it to distributional models, evaluating to what extent the associations found in distributional models can be explained by definition, and how much remains unexplained.

We can use word embeddings to represent distributional models, and we also have resources to represent definitional knowledge – perhaps the largest and most expressive one is WordNet (Fellbaum, 1998), which not only defines over 150,000 words, but also connects the words to each other in relations such as *antonymy*, *hypernymy*, and *meronymy* (Section 3.1). However, the question remains, how can we actually compare word association in these two settings? A number of challenges arise: (1) WordNet is a discrete, graphical structure while word embeddings take the form of continuous real-valued vectors, (2) WordNet defines meaning in the form of *synsets*, or word

senses, which are not directly used in natural language and thus would not appear in word embedding vocabularies, (3) it is difficult to define automatic measures of word association that can be computed for both WordNet and word embeddings, produce reasonable results, and can be compared fairly across settings.

This work addresses each of these challenges. To translate WordNet and word embeddings into common representations, I construct a graph G_W based on WordNet (Section 3) and a graph G_C based on COHA word embeddings (Section 4). In G_W, I include nodes for word lemmas in addition to nodes for synsets, which allows me to compare G_W and G_C on the basis of a sizeable shared lemma vocabulary. I design an approach for operationalizing word association, which utilizes Personalized PageRank to induce semantic fields in each graph, and I define evaluation metrics to quantify the extent to which the definitional results can explain distributional results (Section 5). Finally, I conduct a series of experiments to analyze if there are any properties in the distributional space that can distinguish definitional versus non-definitional association; these experiments provide further insight into how word meaning manifests in distributional models (Section 6).

My main contributions include:

- Developing a novel framework that transforms WordNet and word embeddings into a comparable space.
- Using this framework to quantify the extent to which definitional association explains distributional association (20%-30%), and to discover salient examples of definitional versus non-definitional associations.
- Finding properties in the distributional space that can predict under certain semantic fields whether an association is definitional or not.

These findings are among the first to establish, in quantifiable and interpretable ways, the types of meaning represented in distributional semantics. Furthermore, the utility of the introduced framework, which I have only begun to exploit in this paper, creates the opportunity for many future directions of research. Thus, this work brings us one step closer to disentangling distributional semantics, and to explaining the models that power so many of our state-of-the-art systems.

2 Related Work

The concept of word meaning has been debated extensively in linguistics. Many contrasting camps have formed; for example, semantic externalism, the view that our words and ideas depend on items in the external world (Putnam, 1975), is countered by internalism, the belief that content only depends on properties within our bodies and brains (Segal, 2000). Other debates involve the importance of context (Grice, 1975), or the role of extra-linguistic knowledge, such as common sense (Katz, 1982) or world facts (Fillmore, 1982).

Amid this debate, I posit that there is a place for studying word meaning computationally, as word meaning has already become a part of computational models. The goal of word embeddings such as Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2016), and, most recently, BERT (Devlin et al., 2018) is precisely to represent word meaning so that they can be used for computational tasks, such as automatic summarization, question answering, sentiment analysis, and more. However, as previously discussed, we lack a clear understanding of what types of meaning are encoded in word embeddings, an oversight which may result in potential bias (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018). Thus, the need for this work arises out of the conundrum that we rely on word meaning in our models, yet we lack clarity about the meaning being represented.

My approach is to compare definitional and distributional word association, which prior work has come close to but not quite done. When distributional semantics first became popularized, a common approach was to compare their ability at representing word similarity to the ability of definitional knowledge bases, such as WordNet (Agirre et al., 2009). The models were evaluated based on how well their predictions of similarity matched human judgments of similarity, the gold standard in this scenario, but the models were never evaluated on each other. Thus, my work uniquely flips the framework on its head, by cutting out the third party and directly comparing definitional and distributional models.

Prior research has also examined how to quantify word association in WordNet. Early approaches were based on path length, or variants of it that additionally scaled by the depth of the nodes (Wu and Palmer, 1994) or the maximum

path length (Leacock and Chodorow, 1998). Another family of approaches incorporates information content (Resnik, 1995; Jiang and Conrath, 1997), or makes use of the natural language definitions of the synsets (Pedersen et al., 2004). Hughes and Ramage 2007 show that running random walks over a graph derived from WordNet can produce especially effective measures of word similarity; Gonzalez and Castillo 2012 take this one step further and use random walks to induce groups of associated words, or domains. Hamilton et al. 2016a demonstrate that random walk algorithms can also be used over graphs derived from word embeddings to induce domains. My work synthesizes these findings to run random walks over *both* WordNet and word embedding graphs, and due to the shared graph vocabularies and common method of random walks, I am able to compare the induced domain results and fairly quantify how much they have in common.

3 WordNet: Graph Construction

3.1 WordNet Structure

WordNet (Fellbaum, 1998) is a graph where the nodes are *synsets*, which is a word sense (a meaning) represented as a set of synonymous *lemmas*. A lemma is a base form of a word, e.g. the words "speak," "speaking," and "spoke" share the same lemma, "speak". Each synset is also accompanied by a *gloss*, a definition in the form of natural language. For example, the synset containing the lemmas {"estimable," "good," "honorable," "respectable"} has the gloss, "deserving of esteem and respect."

The edges in WordNet are relations between synsets, with some more common ones being *antonymy* (opposing meanings, e.g. ugly $\leftrightarrow$ beautiful), *hypernymy* (concept to superordinate, e.g. poodle $\rightarrow$ dog), *hyponymy* (concept to subclass, e.g. meal $\rightarrow$ lunch), *meronymy* (has-part, e.g. table $\rightarrow$ leg), and *holonymy* (part-of, e.g. professor $\rightarrow$ faculty). WordNet includes annotations for nouns, verbs, adjectives, and adverbs.

3.2 Constructing G_W

Based on WordNet 3.0[1], I derive a directed graph, G_W. G_W has two types of nodes: (1) *Synset* nodes for every synset, and (2) *LemmaPOS* nodes for every lemma and part-of-speech. For example, "bank-n" and "bank-v" are separate nodes in G_W. I extract almost all WordNet relations to construct bidirectional *Synset-Synset* edges (with weight 1.0). The relations I include are antonym, hypernym/hyponym, meronym/holonym, instance/instance of, entails/entailed by, causes/caused by, attribute/has attribute, derivationally related, pertains to/similar to, and participle. The only relations I exclude are domain relations, which indicate groups of synsets that belong to the same domain, or semantic field. I exclude them because I evaluate G_W based on semantic field induction, so it would be unfair to include annotations of the fields in the graph.

I also construct directed edges from each *LemmaPOS* node to the *Synsets* that it participates in. I weight the edges based on SemCor frequencies[2]; that is, the weight of edge l, s is $p(s|l) = f_{l,s}/f_l$, with +1 smoothing. Intuitively, this encourages a random surfer to move towards more common senses of the $LemmaPOS$. For example, "school-n" maps to several synsets, but its most common one, with the gloss "an educational institution," has weight .813, while a far less common one, glossed as "a large group of fish," only has weight .006. Lastly, I construct directed edges from the *Synset* to *LemmaPOS* nodes, such that a synset s connects to a lemma l if l is one of the synonyms that defines s. As in the former case, I use the SemCore frequencies to define the weight, so the weight of edge s, l is $p(l|s)$.

At the end of the construction process, G_W has 285,912 nodes in total (157,560 lemmas and 128,352 synsets) and 757,110 edges. The graph is well-connected overall, as its largest weakly connected component (WCC) contains 98.8% of the nodes. However, G_W is still quite sparse, with an average unweighted degree of 2.65. This sparsity is beneficial for computational efficiency, and also makes intuitive sense – most lemmas only belong to 1-3 synsets, and most synsets are only connected to a handful of other synsets.

4 Word Embeddings: Graph Construction

4.1 COHA Embeddings

To represent distributional semantics, I utilize word embeddings built by Hamilton et al. 2016b.

[1] https://wordnet.princeton.edu/download/current-version

[2] SemCor is an English corpus annotated for lemmas, POS, synsets, and other features. The SemCor frequencies of (*LemmaPOS*, *Synset*) pairs are included in WordNet.

They release a set of embeddings trained on various corpora using different embedding methods; the version that I am using was trained on the Corpus of Historical American English (COHA) and uses an SVD-based method. I choose COHA because it is representative of English usage – it contains genre-balanced texts for each decade from 1810-2000 – and because it was pre-processed into lemma form, so the embeddings can be directly trained on lemmas instead of raw word forms. This vocabulary of lemmas allows us to directly compare the COHA-based graph with G_W, which also has a large vocabulary of lemmas.

I choose the SVD-based method of embedding because Hamilton et al. 2016a find that this type of embedding is optimal when the embeddings are used to construct a graph and then the graph is utilized for domain induction, which is exactly the goal in this work. The SVD-based method begins by constructing the positive pointwise mutual information (PPMI) matrix of corpus words, such that each entry $\mathrm{M}_{\mathrm{i,j}}^{\mathrm{PPMI}}$ represents the association between vocabulary word w_i and context word c_j, where w_i is said to appear in the context of c_j if w_i occurs within a fixed window of c_j, e.g. within 5 words. The entries of the matrix are defined as:

$$\mathrm{M}_{\mathrm{i,j}}^{\mathrm{PPMI}} = \max\left\{\log\left(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w_i)\hat{p}(c_j)}\right) - \alpha, 0\right\},$$

where $\hat{p}$ represents the smoothed empirical probabilities of the word (co-)occurrences and $\alpha > 0$ provides a smoothing bias that ameliorates the tendency of PPMI to assign extreme scores to rare words (Levy et al., 2015). Clipping the negative values of the matrix has also shown to improve results dramatically; intuitively, this emphasizes positive correlations over negative ones, which are likely to be less meaningful.

The embedding step performs singular value decomposition (SVD) on $\mathrm{M}^{\mathrm{PPMI}}$, such that the word embedding e_i for word w_i is given by:

$$e_i = (\mathbf{U}\mathbf{\Sigma}^{\gamma})_i,$$

where $\mathrm{M}^{\mathrm{PPMI}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$ and $\gamma \in [0, 1]$ is the eigenvalue weighting parameter. SVD embeddings are shown to be more robust than PPMI, since the dimensionality reduction acts as a form of regularization (Hamilton et al., 2016b).

4.2 Constructing G_C

Given the SVD embeddings derived from the COHA 2000s corpus, I then construct the directed graph G_C. I filter the embedding vocabulary to keep only the lemmas that are nouns, verbs, or adjectives, and create a *LemmaPOS* node for each lemma and part-of-speech. For each node, I construct a directed edge to the nodes of the lemma's k-nearest neighbors in embedding space, according to cosine similarity. Following Hamilton et al. 2016a, the weight of edge i, j is also based on the cosine similarity:

$$A_{i,j} = \arccos(-\frac{e_i^{\top} e_j}{||e_i||||e_j||})$$

Another approach to constructing graphs from word embeddings is d-proximity, which connects each word to all words within a fixed distance d. However, prior investigation of embedding networks finds that the k-nearest neighbors approach yields better networks, with greater connectivity, more understandable clusters, and less parameterization (d-proximity needs to adjust d for different embedding types) (Perozzi et al., 2014). Thus, I take the nearest neighbors approach, and with $k = 50$, produce a resulting G_C that has 8,891 nodes and 444,550 edges. The graph is very connected at a global and local level: its largest WCC covers 100% of its nodes, and its average clustering coefficient is .1698. This high[3] degree of clustering is to be expected, since I determined a node's connections by finding its nearest neighbors in embedding space, and geometrically and semantically, two neighbors of the same word are likely to be neighbors of each other as well.

5 Semantic Field Induction

To test how these two graphs model word association, we will draw on the concept of *semantic fields*. Linguists define that two words are similar if they are equivalent or close to equivalent (e.g. "car" and "truck"), but two words can be related without being similar (e.g. "car" and "gasoline"). One of the primary ways that two words can be related is through common semantic fields, where a semantic field is a set of words that are grouped because we understand them to belong to the same topic. Thus, semantic fields are fundamental to word association because they provide the opportunity for words to be associated without

[3] To measure *how* high, we can compare to a configuration model $G_{C,null}$ that has the same degree sequence as G_C but random edges. The average clustering coefficient of $G_{C,null}$ is 0.012, which is 14x smaller than the coefficient for G_C.

being equivalent; for example, "car" and "gasoline" may be associated under the semantic field of *driving*. Furthermore, considering word association through the lens of semantic fields provides a clear and interpretable way of defining relatedness between words. Finally, from a practical standpoint, semantic fields are a useful framework because graph-based models of WordNet and word embeddings have both been successfully used in prior works to induce semantic fields (Gonzalez and Castillo, 2012; Hamilton et al., 2016a), so we may fairly compare semantic field induction across the two settings because they are proven to produce reasonable results in both.

5.1 Methodology

At a high-level, my goal is to ask, given a small set of seed lemmas to represent the core of some semantic field, how does the semantic field spread in the two graphs? In other words, which lemmas do each of the graphs identify as associated with the core set, and to what extent can the associations deduced by the distributional model (G_C) be explained by the associations found by the definitional model (G_W)?

To propagate association from a seed set over the entire graph, I use Personalized PageRank. The PageRank algorithm resolves the equation:

$$r = \beta M r + (1 - \beta)v$$

The first term represents the voting scheme of PageRank, where each page votes on its neighbors and pages with more votes have more importance. Specifically, M is an $N \times N$ matrix, where $M_{ji} = 1/d_i$ if there is an edge from i to j; otherwise, it is 0. The second term represents the probability of teleporting to a random node, and the balance between the two terms is determined by β. In traditional PageRank, the teleport vector has uniform weights, i.e. $1/N$, for all nodes. In Personalized PageRank (PPR), we can determine a specific set of nodes to favor and assign stronger teleport weights to those nodes. The resulting PPR vector spreads the importance of that favored set over the rest of the graph; thus, the computed scores for each node can be seen as the strength of relation between that node and the original set.

With G_W and G_C in hand, we can compare the graphs through their common vocabulary of lemma nodes, which we will call V_L. The size of V_L is 7,773, which covers nearly 90% of the nodes in G_C. Now, given a seed set S of lemmas, I can run PPR in both graphs, and produce association scores relative to S for every node in G_W and G_C. For each graph, I then construct a ranking of the nodes in $V_L - S$, where the node in rank 1 has the highest score, rank 2 has the second highest score, and so on. This set-up allows us to control for the seed set and output vocabulary, so that the crux of experimentation lies in the difference between how the same lemmas connect to one another in the contrasting graphs.

5.2 Evaluation Metrics

To quantify the results, I define two evaluation metrics that represent how well R_W, the ranking of the $V_L - S$ nodes based on G_W, predicts R_C, the ranking of the same nodes based on G_C. First, **Overlap** indicates the proportion of overlap between the first k lemmas in R_W and R_C; this metric can be seen as an extension of hit-rate. The intuition behind cutting off the evaluation at some k is that when we conceptualize a semantic field, we typically only think of our strongest associations, and we do not consider – nor is it meaningful to evaluate – the rankings of words past the top k.

The second metric, $\mathbf{MRR}^+$, is modeled after mean reciprocal rank, which is commonly used to evaluate the correctness of a ranking. Reciprocal rank is defined as $1/r_a$, where a is the first correct answer and r_a is the rank of a; MRR is the average reciprocal rank over a series of queries. However, here we have a single query but k correct answers, since we care about the first k lemmas in R_C. So, I extend reciprocal rank to be:

$$MRR^+_{W,C} = \frac{1}{k} \sum_{a \in A} \frac{r_{C,a}}{r_{W,a}},$$

where A is the set of correct answers. Essentially, this measures the rank ratio, or the degree to which $r_{W,a}$ is worse than the true rank, $r_{C,a}$. We can see that MRR does the same thing, but since it considers only the first correct answer, the true rank is always 1.

Beyond these two summary statistics, I also define two lemma-specific metrics to identify the most salient examples of lemmas for each semantic field. Building off of MRR^+, I am interested in the **"Lean to COHA"** lemmas, i.e. the cases that are strongly associated under G_C but not under G_W. We can identify these examples by sorting the lemmas by their rank ratio, since the ones

that rise to the top must have low r_C and high r_W. To contrast this, I am also interested in **"High in Both"** lemmas that are strongly associated in *both* settings, which we can identify by sorting the lemmas by the maximum of their ranks. In order for them to be early in the ordering, they would need to have low ranks in both settings.

5.3 Experiments

I use the Fast PageRank package[4] to run Personalized PageRank, with β set to 0.85. I distribute the teleport weights uniformly over the seed set, and set the transport weight of the non-seed set nodes to 0. In my evaluations, I set k to 100 for both summary metrics, and I analyze the top 20 "Lean to COHA" and "High in Both" lemmas.

As a baseline, I also build a configuration model of G_W. Since I want to evaluate how well R_W can predict R_C, I compare its performance to the ranking produced by a null model, $G_{W,null}$, with the same nodes and degree sequence as G_W, but randomly swapped edges.

5.4 Results & Discussion

Table 1 presents the results for a range of semantic fields and their seed sets. First, we can see that R_W achieves .20-.32 on the Overlap metric, while $R_{W,null}$ falls around .01-.04. This is striking, when we recall that $G_{W,null}$ still maintains the degree sequence of G_W, which means that it preserves the relative importance of words. What $G_{W,null}$ loses is the connection between words, because it randomly scrambles the edges from G_W. This means that definitional semantic connections improve performance by 5-30x, and that definitional association can explain up to 30% of distributional association.

We see a similar story play out for MRR^+. R_W scores .20-.30 on MRR^+, which we can interpret as, for the top 100 words in R_C, their R_W index is around 4x larger than their R_C index. Considering V_L has 7,773 lemmas, a factor of 4 is decent and represents substantial predictive ability at the true ranking. This also becomes apparent when we compare to the $R_{W,null}$ scores on MRR^+, which range from .03-.07 (a factor of around 20).

Thus, G_W, or definitional association more generally, is able to explain a substantive portion of distributional association, but it is far from explaining all of it. We now turn to the lemma examples to examine where exactly distributional association aligns with definition, and where it goes beyond definition. The High in Both column is mostly what we would expect: words that are inherently associated with the semantic field, such as body parts in the *Body* domain, educational items in *College*, sports and player roles in *Football*, different types of fruit in *Fruit*, ingredients of the courtroom in *Judicial*, and various gendered people in *Woman*.

In contrast, the Lean to COHA column provides associations that require world knowledge, understanding the culture of the semantic field and how it actually manifests in modern human lives. For example, in the *Body* domain, we see exercise-related terms heavily featured, because when we talk about our bodies, it is often in the context of exercise – but exercise is not inherently part of the body, in the way that an elbow or toe are. In the *College* domain, we see disciplines taught in school; in *Football*, we see team names; in *Fruit*, we see ways of preparing fruit; in *Judicial*, we see common (or most talked about) crimes; all of these items are certainly related to their respective semantic fields, but they require world knowledge to recognize the association. Finally, the *Woman* field presents what we might have most expected from non-definitional association: the use of stereotypes and social norms. The words here focus on appearance (e.g. "beautiful," "blonde," "sexy," "graceful"), echoing the disproportionate attention placed on appearance for women in the real world.

This alignment with real-world trends does not end with the *Woman* domain. In fact, the overall dichotomy that we discover between inherent meaning and world knowledge actually maps perfectly onto the Lexicon versus Encyclopedia debate (Peeters, 2000) in linguistics. The Lexicon view is essentially definitional, and "an approach to word meaning can be defined 'encyclopedic' insofar as it characterizes knowledge of worldly facts as the primary constitutive force of word meaning... Our ability to use and interpret the verb 'buy', for example, is closely intertwined with our background knowledge of the social nature of commercial transfer, which involves a seller, a buyer, goods, money, the relation between the money and the goods, and so forth." [5] Linguists have long debated the role of worldly facts in word

[4] https://github.com/asajadi/fast-pagerank

[5] https://plato.stanford.edu/entries/word-meaning/

Seed Set (S)	**Overlap**	$\mathbf{MRR}^+$	**Highest in Both**	**Lean to COHA**
Body: body-n, head-n, neck-n, chest-n, back-n, arm-n, hand-n, leg-n, foot-n	.32 *null*=.02	.301 *null*=.070	elbow-n, toe-n, forearm-n, wrist-n, knee-n, palm-n, fist-n, heel-n, hip-n, finger-n, ankle-n, torso-n, cheek-n, belly-n, breast-n, limb-n, thigh-n, side-n, shin-n, waist-n	dumbbell-n, ache-n, bare-a, gloved-a, flex-n, bend-v, rep-n, curl-n, bruise-n, workout-n, tall-a, shoulder-n, blond-a, rub-v, metabolism-n, prickle-n, rope-n, stare-v, cubic-a, weight-n
College: college-n, degree-n, professor-n, student-n, education-n, university-n	.26 *null*=.02	.240 *null*=.053	faculty-n, school-n, academic-a, educational-a, seminary-n, scholarship-n, teaching-n, enrollment-n, preparation-n, instruction-n, graduate-n, instructional-a, kindergarten-n, high-a, undergraduate-n, schooling-n, master-n, institution-n, teach-v, academy-n	psychology-n, mathematics-n, elementary-a, math-n, sociology-n, semester-n, economics-n, physics-n, psychiatry-n, biology-n, bachelor-n, disability-n, dissertation-n, opportunity-n, campus-n, tuft-n, geography-n, genetics-n, psychologist-n, science-n
Football: football-n, touchdown-n, quarterback-n, lineman-n, coach-n, team-n, game-n	.29 *null*=.01	.279 *null*=.049	linebacker-n, soccer-n, offense-n, play-v, sport-n, receiver-n, playoff-n, tackle-n, hockey-n, defense-n, score-n, player-n, basketball-n, league-n, volleyball-n, scorer-n, rbi-n, bowl-n, coaching-n, tournament-n	season-n, preseason-n, redskin-n, patriot-n, pro-a, falcon-n, raider-n, dolphin-n, viking-n, panther-n, junior-a, stadium-n, texan-n, ram-n, yankee-n, sophomore-n, teammate-n, defensive-a, offensive-a, charger-n
Fruit: pear-n, apple-n, grape-n, banana-n, peach-n, orange-n	.24 *null*=.04	.251 *null*=.057	avocado-n, cherry-n, plum-n, fig-n, lemon-n, fruit-n, almond-n, mango-n, melon-n, pineapple-n, raisin-n, watermelon-n, blackberry-n, raspberry-n, strawberry-n, vine-n, cranberry-n, peel-v, lime-n, celery-n	ripe-a, dessert-n, pie-n, sweet-a, orchard-n, muffin-n, halve-v, pint-n, salad-n, garnish-v, fresh-a, pudding-n, bagel-n, cookie-n, juicy-a, bread-n, prep-n, cream-n, dice-n, mayonnaise-n
Judicial: jury-n, court-n, judge-n, law-n, plaintiff-n, defendant-n, trial-n	.20 *null*=.01	.217 *null*=.034	prosecute-v, justice-n, juror-n, charge-n, prosecution-n, appeal-n, litigation-n, judicial-a, suspect-v, indictment-n, plea-n, courtroom-n, constitution-n, tribunal-n, magistrate-n, legislation-n, authority-n, claim-v, case-n, police-n	federal-a, guilty-a, jail-n, convict-n, murder-n, kidnapping-n, fraud-n, felony-n, legal-a, verdict-n, arrest-n, constitutional-a, allege-v, conviction-n, lawsuit-n, theft-n, rape-n, case-n, homicide-n, robbery-n
Woman: woman-n, girl-n, lady-n, gal-n, feminine-a	.21 *null*=.04	.220 *null*=.075	daughter-n, boy-n, sister-n, man-n, baby-n, gender-n, son-n, girlfriend-n, mother-n, male-a, miss-n, wife-n, female-a, child-n, adult-n, masculine-a, bride-n, sex-n, sexual-a, friend-n	young-a, unmarried-a, beautiful-a, blond-a, blonde-a, tall-a, attractive-a, brunette-n, sexy-a, thank-v, prom-n, little-a, reply-n, aunt-n, whisper-n, shy-a, redhead-n, wish-v, graceful-a, elegant-a

Table 1: Semantic field induction results. The $null$ numbers in **Overlap** and $\mathbf{MRR}^+$ represent the performance of $R_{W,null}$, the ranking produced by $G_{W,null}$. The examples in **Highest in Both** and **Lean to COHA** are the top 20 lemmas in $V_L - S$, when sorted by those metrics.

meaning; it is thus unsurprising the challenges that we face now in computationally modeling word association. However, these results grant us a promising place to start, as they demonstrate that it is possible to take the ambiguity of distributional word associations, quantify how definitional they are, disentangle specific words into distinct categories, and then explain those sets by interpretable patterns that are supported by linguistic theory.

6 Additional Experiments

In the prior section, we discovered that we can use G_W as supervision to disentangle the definitional and non-definitional associations modeled by G_C. Diving in deeper, in this section I conduct a series of short experiments to further compare these types of association, from the perspective of geometric, structural, and temporal properties. The findings illuminate facets of the relationship between definition and world knowledge, and how that relationship physically manifests in distributional representations.

6.1 Is this trivial?

The first question is, is this trivial? That is, I have introduced W_G as a mechanism to separate the distributional associations, but are they already simple to separate due to the geometry of the COHA embeddings? If so, we should be able to recover the labels found by W_G using the embeddings alone and unsupervised clustering. Given the set H, the top 20 Highest in Both words, and the set C, the top 20 Lean to COHA words, we can run a clustering algorithm to group the words into two clusters, then measure the similarity between the predicted clusters and true clusters H and C.

To evaluate the similarity, I consider firstly the accuracy – I treat the problem as a classification task, and take the maximum accuracy when I assign H labels to one predicted cluster and C labels to the other. I also test the adjusted Rand index (ARI), which measures how well the cluster prediction does over a random baseline. The raw Rand index RI is equal to $(a+b)/\binom{N}{2}$, where a is the number of pairs that are in the same cluster in the predicted and true clustering, b is the number of pairs that are in different clusters in the predicted and true clustering, and N is the total number of elements being clustered. ARI adjusts for the expected *RI* of random labeling:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

In my experiments, I use k-means clustering to cluster the embeddings of the H and C words. Since k-means may converge at local optima, I run the algorithm 10 times per experiment, and keep the model with the lowest sum of squared errors. I repeat this process for all of the semantic fields in Table 1; the results are reported in Table 2. We find that for every field besides *College*, unsupervised clustering performs barely above random, which means that H and C do not naturally fall into separable geometric spaces. To further illustrate this point, Figure 1 presents a visualization of the *Judicial* field, where the blue nodes are H words and the red nodes are C words – while the nodes do not seem completely randomly mixed, there is no distinctly red or blue side of the graph. These findings show that the embeddings alone cannot disentangle types of associations, thus confirming the value of an external signal like G_W.

Domain	ARI	Accuracy
Body	.0379	.625
College	.477	.850
Football	.0661	.650
Fruit	-.0183	.525
Judicial	.0663	.650
Woman	.0394	.625

Table 2: Clustering results in all domains.

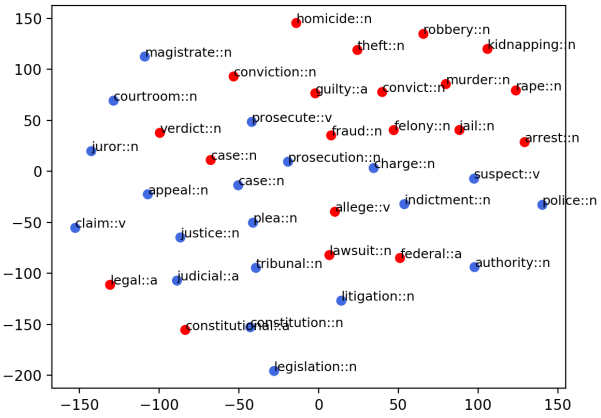

Figure 1: Visualization of the embeddings of the H (blue) and C (red) words in the *Judicial* domain. I use TSNE to reduce the 300-dimension embeddings to 2D.

Domain	In-deg.	Out-deg.	Clust. Coef.	Egonet-in	Egonet-out	PageRank
Body	(.118)	(-.059)	-.239	(-.015)	(.105)	(-.121)
College	(-.135)	(.067)	.369*	(-.074)	-.359*	(-.086)
Football	-.517**	-.349*	.485**	-.428**	-.586**	-.539**
Fruit	.401*	(.118)	-.336*	(.146)	.460**	.310
Judicial	.411**	.507**	.304	.479**	(.154)	.443**
Woman	(-.103)	(-.199)	(-.042)	(-.130)	(.001)	(-.025)

Table 3: Pearson correlation between structural properties and Lean to COHA label. ** indicates p $<$ .01, * indicates p $<$ .05, and parentheses around the number indicates p $>$ .20.

6.2 Structural properties

In this section, I explore the structural properties of the H and C nodes in G_C, and test whether there is any correlation between a certain property and belonging to C. The properties that I consider are the weighted in-degree (**In-deg.**), weighted out-degree (**Out-deg.**), clustering coefficient (**Clust. Coef.**), number of edges in the node's egonet (**Egonet-in**), number of edges connecting the node's egonet to the rest of the graph (**Egonet-out**), and PageRank (**PageRank**).

Table 3 presents the results: interestingly, we find that each property is significantly correlated with belonging to C, but only for some of the semantic fields and often in opposite directions. For example, PageRank is negatively correlated for *Football*, but positively correlated for *Judicial*, which we can interpret as suggesting that in *Football*, terms become less central when we move from definition to world knowledge, while in *Judicial*, terms become more central. This is reasonable given what we saw of the actual lemma examples: the Lean to COHA terms for *Football* were specific team names, while for *Judicial* they were common crimes. For the rest of the properties, in-degree and out-degree show similar results to PageRank, but clustering coefficient and the egonet features differ in which sets of semantic fields they can predict.

One takeaway from these statistics is that we have further shown that the findings in Table 1 are not trivial. The graph-based approach cannot be reduced to simple artifacts of the graph, but rather seems to reflect something deeper about the semantics being modeled. It is particularly fascinating that there are significantly predictive properties for four of the semantic fields, but those properties are not consistently predictive. This suggests, perhaps, that within a semantic field, there is a consistent transformation to be found from definitional to worldly space, but across semantic fields, the transformation does not hold.

6.3 Temporal Properties

My final question is whether there are temporal properties of distributional associations that might be able to distinguish definition versus world knowledge. It would be reasonable to hypothesize that the world knowledge changes more quickly over time, because it is at the whims of culture and trends, while the former is grounded by inherent meaning. To test this hypothesis, I leverage a century of COHA embeddings, and test whether the most "persistent" associations correlate with the most definitional.

First, I build a graph, $G_{C,j}$, for every decade j in the 1900s, following the graph construction process described in Section 3.2. Then, for each semantic field, I run Personalized PageRank to induce that field in every $G_{C,j}$; see Section 5.1 for details on semantic field induction. Now, for every word i, I have a vector v_i of length 11 (1900s-2000s) representing i's level of association with a given field for every decade. Our goal is to identify the words that have been consistently, highly associated with the semantic field over time – thus, we want the most "persistent" words, or the ones with the largest area-under-curve (AUC) for v_i. To visualize this concept, consider the curves in Figure 2 of several words with respect to the semantic field *Body*: "dumbbell" has only been associated with *Body* in the last decade, which is why its curve shoots up from 1990 to 2000, while "wrist" has been highly associated with *Body* across all decades, and "weight" falls in between them.

Now, I compare words in H to words in C, and test whether $AUC(v_i)$ is correlated with the H label. Just like for the previous properties, I find that this temporal property is correlated but only for some of the semantic fields. For *Body*,

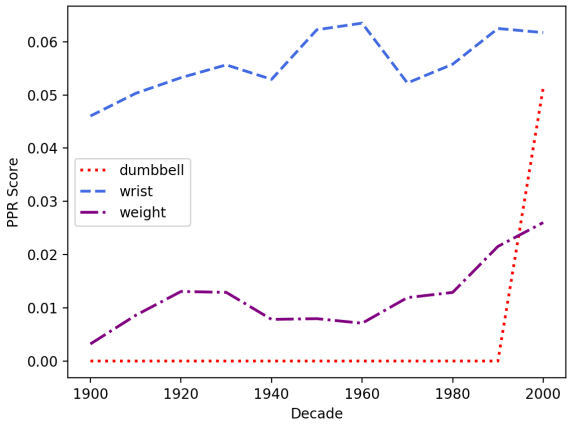

Figure 2: Temporal COHA association vectors for three nouns with respect to the semantic field *Body*.

the correlation is very strong, with $r = .586$ and $p < 10^{-4}$. *Football* and *Woman* are weakly correlated ($p < .08$), but the rest of the semantic fields do not yield anything significant. We might hypothesize that these three fields are relatively more reliant on temporal trends, with norms about physicality and gender rapidly changing, and football being caught up in current players and teams, but we would need more rigorous studies to back up these claims.

What we can say with more certainty is that, again, a property has predictive power but only for a subset of the fields. Furthermore, note that all of the semantic fields are now accounted for by at least one property: *Body* and *Woman* associations can be distinguished by temporal features, *College* by clustering, *Fruit* and *Judicial* by structural properties, and *Football* by structural and temporal. This strengthens the notion that there is a strong relationship between definitional and non-definitional meaning that is reflected physically in distributional space, but that relationship is not universal across semantic fields – which makes it all the more intriguing.

7 Conclusion

In this work, I have laid out a daunting problem space and begun to address its challenges. Disentangling distributional models is necessary, as they form the backbone of many of our systems, yet their current lack of transparency leads models astray and often in undesirable directions. However, even the initial step of disentangling definitional versus non-definitional association using WordNet is fraught with difficulty: how to project WordNet and word embeddings into a common representational space, how to compare their differing vocabularies, how to learn word association from their respective graphs, and how to quantify a comparison of their results.

The framework I introduce of parallel graphs, derived from differing sources but common in their lemma vocabulary, allows for such comparison. My methods to induce semantic fields yield reasonable results from both graphs, and my evaluation metrics summarize the distance between results and discover salient examples that can be used for fine-grained analysis. The final set of experiments confirm that the framework discovers valuable information that could not have been found from word embeddings alone, but also hints that embeddings do encode some signals that can be exploited to learn the relationship between definitional and non-definitional association.

This work also has its limitations, one of which is that WordNet is imperfect. For example, it is more detailed in certain areas than others, and its annotation is better for nouns than other parts-of-speech. Future work should evaluate the associations represented by WordNet against other dictionaries or taxonomies. Another limitation is evaluation: I have shown that the induced semantic fields seem reasonable and based my results on only the top k words, all of which I analyzed and reported, but more rigorous evaluation should be conducted, such as showing the predicted associations to human annotators. One final future direction is further removed from the current work, but builds on the results of the final experiments. If we could use the WordNet signal to learn how to classify definitional meaning in embeddings, we could hope to transfer this information into new embedding spaces. The ultimate goal could be to automatically generate dictionaries from natural language alone, which would be valuable for any application lacking a dictionary, such as non-standard domains or low-resource languages.

8 Acknowledgements

I would like to thank Dan Jurafsky and Jure Leskovec for their helpful comments on this work. Thank you also to the CS224W teaching assistants for their time and hard work, and to Jure and Michele Catasta for teaching a wonderful course.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL-HLT*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*. Hanshin Publishing, Seoul.

John R. Firth. 1957. *Studies in linguistic analysis, 1-32*. Oxford: Blackwell.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 155(16).

Aitor Gonzalez and Mauro Castillo. 2012. A graph-based method to improve wordnet domains. In *CICLING*.

Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016a. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL*.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *EMNLP*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING*.

Jerrold J. Katz. 1982. Common sense in semantics. *Notre Dame J. Formal Logic*, 23(2).

Claudia Leacock Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 265–283. MIT Press.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. In *ACL*.

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

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *NAACL-HLT*.

Bert Peeters. 2000. Setting the scene: recent milestones in the lexicon-encyclopedia debate. In Bert Peeters, editor, *The lexicon - encyclopedia interface*. Elsevier Science, Oxford.

Bryan Perozzi, Rami Al-Rfou, Vivek Kulkarni, and Steven Skiena. 2014. Inducing language networks from continuous space word representations. In *Studies in Computational Intelligence*.

Hilary Putnam. 1975. The meaning of 'meaning'. *Language, mind, and knowledge*, 7.

Phillip Resnik. 1995. Using information content to evaluate semantic similarity in a taxanomy. In *IJCAI*.

Gabriel Segal. 2000. *A Slim Book about Narrow Content*. Cambridge Mass: MIT Press.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *ACL*.