Relating semantic similarity and semantic association to how humans label other people

Kenneth Joseph
Northeastern University
k.joseph@northeastern.edu

Kathleen M. Carley
Carnegie Mellon University
kathleen.carley@cs.cmu.edu

Abstract

Computational linguists have long relied on a distinction between semantic similarity and semantic association to explain and evaluate what is being learned by NLP models. In the present work, we take these same concepts and explore how they apply to an entirely different question - how individuals label other people. Leveraging survey data made public by NLP researchers, we develop our own survey to connect semantic similarity and semantic association to the process by which humans label other people. The result is a set of insights applicable to how we think of semantic similarity as NLP researchers and a new way of leveraging NLP models of semantic similarity and association as researchers of social science.

1 Introduction

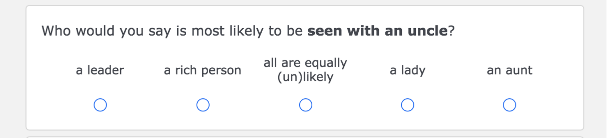
Computational linguists often find it useful to distinguish between the *semantic similarity* and *semantic association* of two concepts (Resnik, 1999; Agirre et al., 2009; Hill et al., 2016). Two concepts are highly semantically *associated* if when we think of one, we almost always think of the other. In contrast, two concepts are semantically *similar* if they share some salient property. Resnik (1999) differentiates between similarity and association via the following example: "cars and gasoline [are] more closely [associated] than, say, cars and bicycles, but the latter pair are certainly more similar".

This distinction between semantic similarity and semantic association is important to computational linguists for two reasons. First, different types of models are engineered to infer one versus the other (Sahlgren, 2006). For example, topic models (Blei et al., 2003) are geared towards inferring sets of semantically associated concepts, while neural embedding models (Mikolov et al., 2013; Levy and Goldberg, 2014; Zhang et al., 2013) aim to place concepts into a latent space where proximity indicates semantic similarity. Second, distinguishing between semantic similarity and semantic association can help us understand how well these models are optimizing for their intended purpose. For example, Hill et al. (2016) develop a dataset of semantic associations and similarities measured via survey which is used to show that many neural embedding models are actually much better at capturing association than they are at capturing similarity.

The present work uses these survey-based measurements from Hill et al. (2016) to better understand an entirely different question - what is the process by which individuals label other people? Specifically, we focus on understanding how semantic associations and similarities between *identities*, defined as the labels that we apply to people (e.g. man, woman, etc.) (Smith-Lovin, 2007), impact this labeling process. We focus here on the following two hypotheses:

- *H1:* The higher the semantic similarity between two identities, the more likely two identities are to be applied to the same person (e.g. this person is both a woman and a scholar)
- *H2:* The higher the semantic association between two identities, the more likely two identities are to be applied to two people in the same

Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 1–10,
Austin, TX, November 5, 2016. ©2016 Association for Computational Linguistics



Who would you say is most likely to be **seen with an uncle**?

a leader	a rich person	all are equally (un)likely	a lady	an aunt
○	○	○	○	○

Figure 1: An example of a "SeenWith" question as seen by participants

context (e.g. a doctor is often seen with her patient)

As a canonical question in the social sciences, a significant amount of work has studied the way people label others. Social psychologists have studied both how we label ourselves (Stryker and Burke, 2000; Owens et al., 2010) and how we label others (Heise, 2007; Penner and Saperstein, 2013), as have cognitive psychologists (Kunda and Thagard, 1996; Freeman and Ambady, 2011), neuroscientists (Cikara and Van Bavel, 2014) and linguists (Recasens et al., 2011; Bucholtz and Hall, 2005). Despite the depth and breadth of this work, however, few quantitative models exist that can actually *predict* how an individual will be labeled in a particular situation. Where such models do exist, they tend to either focus explicitly on similarity *or* association (Joseph et al., 2017), to conflate the two and treat them both as semantic "links" in cognitive networks (Freeman and Ambady, 2011), or to ignore relationships between identities all together in favor of feature-based models of individual identities (Heise, 2007).

By testing the two hypotheses above, the present work hopes to achieve three related goals that further our understanding of the identity labeling process. First, we would like to provide additional evidence that rather than focusing simply properties of identities in isolation, we must incorporate identity relationships into our models of how people label other people (Kang and Bodenhausen, 2015; Joseph et al., 2017). Second, we hope to provide evidence that it is not merely enough to consider relationships between identities - if our hypotheses are correct, they would indicate that different types of relationships impact how we label others in distinct ways. Finally we hope to show that differentiating similarity from association is a useful and parsimonious way to characterize these different types of relationships.

In the following sections, we describe the data from Hill et al. (2016) that we use as measurements of semantic associations and semantic similarities. We then detail the development of a survey, intended to test the two hypotheses above, that asks respondents to label people in hypothetical social situations. The survey asks respondents to perform identity labeling by providing answers to multiple choice questions, an example of which is given in Figure 1 for one of the many identities (uncle) we consider here.

In addition to asking questions of the form "who would you say is most likely to be *seen with* an uncle?", as shown in Figure 1, we also ask questions of the form "given that someone is an uncle, what other identity is *that same person* most likely to also be?" These two different types of questions get at H1 (above) and H2 (Figure 1). Even intuitively, we can see that they should have different mechanisms by which individuals determine the appropriate label. In the first question above, for example, people would be likely to respond with "aunt". However, this is among the least likely answers to be given in the second question, as "uncle" and "aunt" are mutually exclusive role identities. While these questions shrink the complex process by which identity labeling occurs down to a simple survey, they therefore are useful as a starting point for exploring the importance of semantic similarity and semantic association in the identity labeling process.

2 Data

For this study, we use a set of 88 pairs of identity words for which data on semantic similarity and semantic association scores already exists. These scores are drawn from the SimLex-999 dataset of Hill et al. (2016), which includes survey measurements of both semantic association and semantic similarity for 999 pairs of concepts. For the purposes of the present work, we were only interested in concept pairs from the SimLex-999 dataset in which both concepts were unambiguously identities, thus the reduction to only 88 pairs of words.[1]

To measure semantic association, Hill et al.

[1] We did not consider the pair heroine-hero, as it appeared that the former term was interpreted as the drug rather than the female hero. We also ignored the terms god, devil and demon, judging them to be more representative of the religious concepts than their alternative identity meanings

(2016) used the USF free association dataset compiled by Nelson et al. (2004). This dataset contains five thousand "cue" words that were given to at least ninety-four survey respondents (mean = 148). For each cue, respondents were asked to write the first word that came to mind that they thought of when shown the cue. As a result, for each cue word one can construct a distribution of its association to other words based on the percentage of survey respondents that gave that word as an answer.

For a survey-based measure of semantic similarity, Hill et al. (2016) pulled 900 of the 72,000 possible pairs of cue-association words from the USF Free Association dataset. To this dataset, they add 99 pairs of words found in the USF Free Association dataset where each was either a cue word or a response word but that were not themselves associated. For each of these 999 pairs of concepts, the authors then asked approximately 50 respondents on Amazon's Mechanical Turk to rate the similarity of each pair of concepts. They used a scale defined via examples similar to the one from Resnik (1999) presented above and allowed respondents to compare the similarity of concept pairs. Additionally, it should be noted that Hill et al. (2016) assume semantic similarity is symmetric, but do not directly test this point.

Table 1 presents some examples of the 88 identity pairs we extracted from the SimLex-999 data based on whether they were higher or lower than average on one or both dimensions. Broadly, we see that identities which are highly similar seem to be those one might be willing to apply to the same individual, and identities that are highly associated are those one might tend to see together in similar social contexts. These obvious examples suggest support for our hypotheses - we now detail the survey we develop in order to more formally test these intuitions.

3 Identity Labeling Survey Description

Let us assume two identities A and B make up one of the 88 pairs of identities we draw from the SimLex-999 dataset. To construct our surveys, we first generated eighty randomized questions with this pair, twenty each from four types:

- **"IsA" A questions:** "Given that someone is a[n] **A**, what is **that same person** most likely to also be?"
- **"IsA" B questions:** "Given that someone is a[n] **B**, what is **that same person** most likely to also be?"
- **"SeenWith" A questions**: "Who would you say is most likely to be **seen with** a[n] **A**?
- **"SeenWith" B questions**: "Who would you say is most likely to be **seen with** a[n] **B**?

Each of these questions had five multiple choice answers. Within the answer set, the identity not in the question itself (i.e. B if A was in the question, or vice versa) was given as one of the answers. As shown in Figure 1, we then included three random identities from a set of 234 commonly occurring identities[2] as alternative choices, along with the option "all answers are equally (un)likely" in order to allow respondents to opt out of answering questions they were uncomfortable with.

These questions were then distributed as surveys where each respondent saw 40 random questions. With 80*88=7,040 questions to ask, we therefore required 176 respondents. Surveys were deployed on Amazon's Mechanical Turk to only "Masters"[3] and only those with IP addresses within the United States. To assess accuracy for respondents, we randomly sampled 5 questions from each respondent and ensured that answers appeared reasonable. No personally-identifiable information was collected, and all (anonymized) survey questions, responses and analyses presented here are available at https://github.com/kennyjoseph/nlp_css_workshop.

4 Results

As we will show in this section, our results show support for both hypotheses. High similarity between identities led to more 'IsA' attributions (H1),

[2] Due to space constraints, how these identities were chosen is not described here - for more details, we refer the reader to Section 5.4.2 of (Joseph, 2016)

[3] https://www.mturk.com/mturk/help?helpPage=worker#what_is_master_worker

Similarity, Association	**Examples**
High Similarity, High Association	physician & doctor, friend & buddy; student & pupil; teacher & instructor
High Similarity, Low Association	buddy & companion; adversary & opponent; author & creator; champion & winner; leader & manager; politician & president
Low Similarity, High Association	wife & husband; woman & man; child & adult
Low Similarity, Low Association	adult & baby; author & reader; boy & partner; chief & mayor; dad & mother; daughter & kid; friend & guy; girl & maid; guy; & partner; king & princess; lawyer & banker

Table 1: Examples of identities that are higher or lower than average for each combination of high/low of semantic similarity and semantic association.

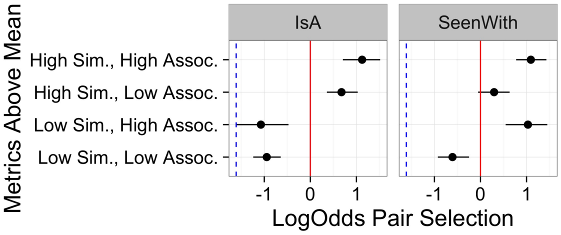


Figure 2: On the x-axis, the log odds of selection. On the y-axis, identity pairs are split into the same categories as in Table 1; see text for details. For each category, 95% bootstrapped Confidence Intervals are presented for mean log odds of selection of all identity pairs within the category. Vertical lines are drawn at a log-odds of selection of 0 (red solid line; 50-50 chance of selection) and at $log(\frac{1}{5})$ (blue dashed line; random chance of selection)

while high association led to more 'SeenWith' attributions (H2).

Figure 2 presents a high-level overview of the results in terms of the classification of high/low similarity/association presented in Table 1. Similarly to Table 1, the y-axis of Figure 2 presents four "categories" of identity pairs based on whether they were above ("high") or below ("low") average on the two different semantic metrics.[4] The x-axis of Figure 2 shows a 95% confidence interval for the *log-odds of selection* of all identity pairs in the given category. The *log-odds of selection* for an identity pair is the (log) proportion of times an identity pair element in the answer set was selected out of the 20 randomized questions generated for that question type and that arrangement of identities. So, for example, if "woman" were selected 19 out of 20 times when given as a possible answer for the question "Who would you say is most likely to be seen with a man?", then the log-odds of selection for the "man-woman" pair for "SeenWith" questions would be $\frac{19+1}{1+1}$, where a $+1$ is added to avoid zero-valued denominators. Finally, Figure 2 also displays two baselines to consider- a red, solid line is drawn at a log-odds of 0, representing the odds of the identity being selected as the answer more than 50% of the time. The blue, dashed line is drawn at a log-odds of 20%, that is, the odds of the identity being selected more often than random.

Figure 2 provides evidence that high semantic similarity breeds high log-odds of selection for "IsA" questions (H1), and high association breeds high log-odds of selection for "SeenWith" questions (H2). However, two anomalies not suggested by our hypotheses are worth considering as well. First, note that when both similarity and association are low, the log-odds of selection are still noticeably greater than chance. This is likely due to the way that word pairings were selected in the SimLex-999 dataset- Hill et al. (2016) sampled largely from existing cue/response pairs in the USF free association data. Consequently, we work here with identity pairings that already have some form of association in at least one direction; their relationship is therefore stronger than a random baseline in almost all cases. Second, we see that semantic similarity appears to have a strong impact on "SeenWith" questions - that is, identities which are above average in semantic similarity but *not* on semantic association still are perceived to frequently exist together in the same context.

[4] Note that Table 1 shows only some examples of each category, whereas Figure 2 uses the entire dataset

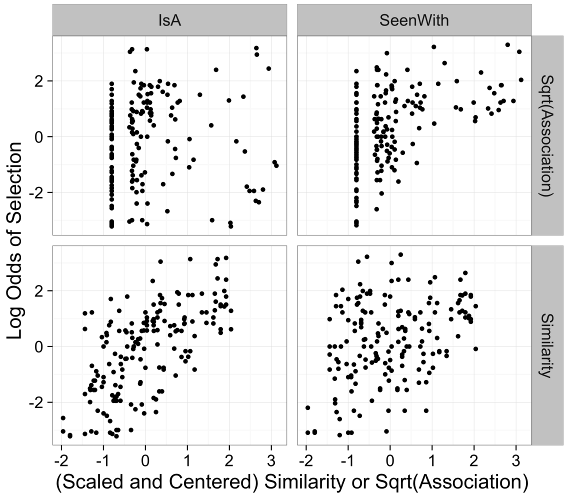
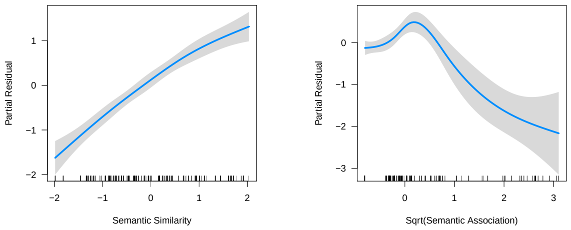
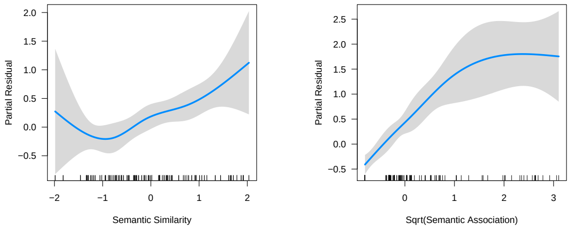


Figure 3: A scatterplot of bivariate relationships between the two dependent variables and the independent variable. Each point represents one identity pair. Results for association for IsA questions (top left), association for SeenWith questions (top right), similarity for SeenWith questions (bottom right) and similarity for IsA questions (bottom left) are presented



(a) (b)

Figure 4: Results from a GAM fit to logit of the odds of selection for "IsA" questions. Figures a) and b) show fit lines (blue bar) and 95% confidence intervals of the fit (grey shadows) for semantic similarity and semantic association, respectively.



(a) (b)

Figure 5: The same GAM model as in Figure 4, except here we fit to data from only "SeenWith" questions

These observations are also supported by Figure 3, which portrays four scatterplots of the bivariate relationships between similarity and the square root of association[5] for both IsA and SeenWith questions. However, because similarity and association are themselves related, it is important to leverage a more rigorous statistical approach that allows us to see the relationship between one of our factors (similarity/association) while controlling for variance in the other. We fit a binomial generalized additive model (GAM) (Hastie and Tibshirani, 1990) using the `mgcv` package in R (Wood, 2006; R Core Team, 2015) to results on each type of question independently.[6] In essence, generalized additive models are generalized linear models that relax the assumption of linearity, instead fitting a (possibly multi-dimensional) curve to each model parameter. The "wigglyness" of these curves, or functions, is controlled by some form of penalty; in the `mgcv` package, this penalty is determined via cross-validation.

The model we use predicts the logit of the odds of selection by fitting tensor product bases to the (scaled and centered) measures of semantic similarity and the square-root of semantic association independently as well as a multivariate tensor basis on their interaction. Figure 4a) shows the fit on semantic similarity to partial residuals of the logit odds of selection for IsA questions only. Figure 4b) shows the same for (the square root of) semantic association. Partial residuals essentially show the fit of each variable after "removing" effects of the other variable and their interaction.

Figure 4a) shows that, controlling for association and the interaction effects of the two variables, semantic similarity has a strong, positive effect on the log-odds of selection in IsA questions. This result provides further support for *H1*. Interestingly, however, we see in Figure 4b) that there exists a sort of acceptable region of association for "IsA" questions.

[5] We use the square root as it better represents a strong distinction between a zero-valued association and a small but nonzero association. We feel this is conceptually appropriate, as a difference between any association and no association seems more important than a difference between some association and more association. Importantly, however, results presented here are robust to this decision and also robust to removing zero-association pairs all together, see the replication material for these robustness checks.

[6] Chapter 8 and Chapter 9 of (Shalizi, 2013) provide a nice introduction to GAMs and tensor product bases.

Association increases the log odds of selection up until a point but then shows, net of similarity, a significant *negative* effect on the odds that survey respondents would label the same person with those two identities. The marginal positive relationship, which holds even when we remove zero-association identity pairs, is interesting but appears to be related to oddities with how association is measured by Hill et al. that we discuss below. On the other hand, as we will discuss in Section 5, the eventual negative effect of association on "IsA" questions seems to be largely attributable to the existence of role/counter-role pairs, such as "uncle/aunt" and "husband/wife". These relationships have been heavily theorized but have been notoriously difficult to measure (Burke, 1980), thus our finding presents a novel quantification of an important and well-known phenomena.

Figure 5 provides the same model except fit on data from SeenWith question responses. Here, we observe that semantic association has, as expected in *H2*, a strong impact on log-odds of selection. We also see that net of semantic association and the interaction term, semantic similarity still has a significant positive effect on log-odds of selection.

Figure 4 and Figure 5 thus provide confirmation of *H1* and *H2*, as well as providing two novel insights. First we see that even when identities are semantically disassociated, underlying similarity (e.g. in the case of the identity pairing adversary and opponent) can impact our perception of which identities are likely to be seen together. Second, we see that high levels of semantic association can actually decrease the odds that two labels will be applied to the same individual. This further emphasizes the need to characterize similarity and association as distinct measures in models of the identity labeling process.

Before following up on these two points, we further note that Figure 4 and Figure 5 show large standard errors for the fit lines (particularly at the ends of the distribution), suggesting the models struggled with outliers. Table 2 shows the ten cases in which the "SeenWith" model most heavily under-predicted the true log-odds of selection. The table presents some surprising scores for semantic association - for example, "king" and "princess", as well as "king" and "prince", are both less associated than the average identity pair in our dataset. Given that these identities are drawn from a similar domain, these numbers are surprising at first sight.

The cause of this is, we believe, the use of the proportion of free association connecting two concepts by Hill et al. (2016) (and others) as a metric for semantic association. The problem with using this metric is that a single association can "eat up" a significant amount of the semantic association in a free association task, masking minor but still important associations. Specific to the case of "king", the identity "queen" takes most of the association score in a free association task, meaning other identities that are still highly associated are given lower scores than we might expect. A related example is the identity mother, which has a high free association score to "father" but no association with "dad".

Predictions for our "SeenWith" model are thus hindered by the specific way in which semantic association is measured. The same can be said for the results of the "IsA" model - more specifically, the measurement assumption of Hill et al. (2016) that semantic similarity is symmetric leads to difficulties in prediction. Table 3 presents ten identity pairs where log-odds of selection differed the most depending on which identity was presented in the question. As pairs had the same value for semantic similarity regardless of which identity was presented first, these pairs represent obvious cases where the model would be unable to capture variance in the data. They also present obvious cases where semantic similarity cannot be assumed to be symmetric. For example, a "president" tends to be a "politician", but a politician is not always a president. These asymmetries are due to the naturally occurring hierarchy of identities, and emphasize the variety of ways in which identities can be considered to be similar.

5 Discussion

Results from our survey can be summarized as follows:

1. *H1*- that higher semantic similarity would increase the likelihood two identities are to be applied to the same person, and *H2* - that higher semantic association would increase the likelihood two identities are to be applied to two people in the same context - were supported

Rank	Identity Given in Question Text	Identity Given as Possible Answer	Pred. Log-odds (from GAM)	Actual Log-odds of Selection	Scaled Semantic Association (sqrt)
1	captain	sailor	-0.40	2.35	-0.80
2	sailor	captain	0.33	3.00	-0.08
3	author	reader	-1.28	0.98	-0.80
4	worker	employer	0.23	2.40	-0.32
5	king	princess	-0.38	1.79	-0.80
6	princess	king	0.09	2.23	-0.10
7	king	prince	0.30	2.40	-0.28
8	employee	employer	1.14	3.22	1.03
9	professor	student	-0.13	1.85	-0.31
10	president	politician	0.52	2.48	-0.32

Table 2: Top ten identity pairs for the "SeenWith" model in terms of under-prediction by the model relative to the true log-odds of selection by survey respondents. "Identity Given in Question Text" is the identity presented in the survey question, i.e. the A in "Seen With" A questions above; "Identity Given as Possible Answer" would then be the B. Semantic association is mean-centered and scaled by 1SD.

	Identity 1 (ID1)	Identity 2 (ID2)	Log-odds, ID1 first	Log-odds, ID2 first
1	stud	guy	-1.61	1.73
2	president	politician	-0.17	3.14
3	princess	bride	-2.48	0.41
4	worker	employer	0.98	-1.85
5	warrior	man	-2.08	0.56
6	teacher	rabbi	0.15	-2.25
7	mayor	chief	-0.08	-2.40
8	manager	leader	-0.37	1.85
9	baby	adult	-3.09	-0.89
10	worker	mechanic	1.22	-0.76

Table 3: Top ten identity pairs in terms of difference in log-odds of selection in "IsA" questions depending on which identity was presented in the question (vs. as a possible answer)

2. High semantic similarity is indicative of high log-odds of selection for "SeenWith" questions

3. Semantic association has a curvilinear impact on "IsA" questions - after some point, high semantic association between identities translates to lower odds of selection

4. Limitations of the measurement model of Hill et al. for semantic similarity (assumption of symmetry) and semantic association (proportional measurement model) in our context breed interesting outliers

Support for *H1* and *H2* was shown in both exploratory analyses and more rigorous statistical modeling of the data. Of more interest in this section, however, are the latter three points, which we feel require some further discussion.

With respect to the fourth point above, our results suggest that evaluations using Hill et al.'s (2016) data may be better served by making two additional assumptions. First, we suggest a stricter adherence to Tversky's theory of semantic similarity (Tversky and Gati, 1978), which argues that symmetry cannot be assumed in measurements of the similarity between two concepts. Second, we suggest that alternative measurements of semantic association, such as those based on spreading activation (Collins and Loftus, 1975), may be better representations of semantic association than a simple proportion based on survey responses.

With respect to the second point above, similarity's positive impact on "SeenWith" questions, we believe this to be indicative of an important tension in the linguistic definition of semantic similarity by, e.g., Resnick (1999) and the way we apply multiple identities to the same individual. This is because two distinct forms of similarity seem to play a role in how respondents answered questions. Similarity as typically defined, and thus measured, by computational linguists tends to represent taxonomic relationships, as in, "a lawyer `isA` professional". However, with respect to identity, similarity also refers to labels that may apply to the same individual regardless of taxonomic relationship - in sociological terms, the extent to which two identities are cross-

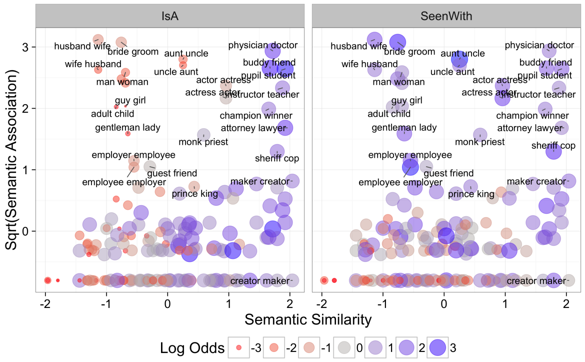


Figure 6: Results for the two different types of questions for log-odds (represented by point color and size), semantic association and semantic similarity. Within each subplot, each identity pair is shown by two points, one each depending on which identity was shown in the question and which was given as a possible answer. Outlier points are labeled based on low-probability with an overlying density estimator

cutting (Blau, 1977). Cross-cutting similarities are unlikely to be captured via denotatively organized data sets like WordNet, or even, it seems, from explicit questions about semantic similarity.

Where they do seem to be captured, however, is in the survey methods presented in this article. A good example is the identity pair "secretary-woman", which had a log-odds of selection of 1.22 (sixteen out of twenty) and a scaled semantic similarity of -1.29 (1.29 standard deviations below the mean). These two identities have little, if any, denotive similarity relationship, and it seems that when the question of similarity is posed explicitly as by Hill et al. (2016), respondents were focused on this connection.[7] In contrast, via the more subtle mechanisms utilized used in our survey, we see the well-known gender bias towards considering secretaries as being primarily women. An important question for NLP+CSS is how to understand and model these subconscious, culturally-based similarities as contrasted with the more traditional taxonomical notions of similarity, and interesting work has certainly begun along these lines (van Miltenburg, 2016; Beukeboom and others, 2014; Bolukbasi et al., 2016).

Finally, Figure 6 provides some insight into the third point above, the negative relationship between semantic association and "IsA" questions. In the figure, we show two subplots, one each for the two different types of questions. Within each subplot, each of the 88 identity pairs studied is given by two points, one each depending on which identity was shown in the question and which was given as a possible answer. The x-axis displays the (scaled) semantic similarity of the identity pair, the y-axis displays the (scaled) square root of semantic association.[8] Finally, each point is colored and sized in Figure 6 by the log-odds of selection - the darker the blue and larger the point, the higher the log-odds, the darker the red and smaller the point, the lower the log-odds.

Figure 6 shows that identities high in association but low in similarity do indeed have very low log odds of selection in "Is-A" questions. Looking at the labels of these identity pairs, we see that they tend to be, intuitively at least, in direct opposition to each other - e.g. husband and wife, man and woman, etc. A more restricted class of such opposing identity pairs, those that fill opposing *roles*, are referenced in classic quantitative models of identity as *role/counter-role* pairs (Burke, 1980). We observe a broader class of *identity/counter-identity* pairs in Figure 6 which are easily discerned by contrasting their semantic association with their semantic similarity.

While many identity/counter-identity pairings are intuitive and have long been studied by social scientists, to the best of our knowledge no methods currently exist to automatically enumerate such pairs. Antonym definitions in lexical databases like WordNet would seem to be one useful resource for this task, but are missing several of what we consider to be basic identity/counter-identity pairings (e.g. groom/bride). Our work also certainly does not fit this bill of automated methods, as we use data curated by survey. Thus, as NLP tools develop to better infer semantic similarity, uncovering identity/counter-identity pairings is one useful

[7] This extends to lexical databases like WordNet as well, where there is no obvious taxonomical connection between these concepts

[8] Note that several zero-associations in Figure 6 are the result of our use of both "directions" of each identity pair. Thus, while we are guaranteed some non-zero association in most of the pairs collected by Hill et al. (2016) in one "direction", in the other there is no such guarantee.

application. While observing intuitive pairings, e.g. man-woman, may not be particularly interesting, extracting less intuitive identity/counter-identity relationships from text, for example, those marking opposing ethnic factions, is a very important avenue of application for these models.

6 Conclusion

In the present work, we leverage measurements and conceptualizations of semantic similarity and semantic association by NLP researchers to study how individuals label other people, a canonical problem in sociology and social psychology. We find strong support for our hypotheses that semantic similarity is related to which identities we choose to apply to the same individual and that semantic association is related to which identities we choose to apply to different individuals in the same context.

Beyond confirmation of these hypotheses, our work presents several other useful contributions of use to the fields of NLP and Computational Social Science (CSS). With respect to NLP, an analysis of outliers in our results suggests that Hill et al.'s (2016) measurements, commonly used to evaluate neural embedding models, may have important restrictions not previously noted by the community. Thus our results suggest that the way people label others provides unique insights into measurements of similarity and association beyond those currently explored by common NLP evaluation datasets.

With respect to CSS, we have given evidence that identity relationships are important in understanding the identity labeling process, that there are unique types of these relationships with disparate impacts on this process, and that similarity and association are a powerful yet parsimonious means of characterizing these types of relationships. In addition, we find that differentiating identities by their semantic associations and semantic similarities provides an interesting socio-theoretic definition of identity/counter-identity pairs, a classic relational model of identity measurement (Burke, 1980). Our work therefore suggests new directions for theoretical development in CSS beyond just the way we label others. As we move towards better understandings of and better models of extracting semantic similarity from text, we see this as an exciting avenue of future work at the intersection of NLP and CSS.

Future work should also serve to address the limitations of the efforts presented here. In particular, the bias in using these particular 88 identity pairs from the SimLex-999 dataset is unclear. Further, social scientists also often assume that both affective meaning of identities and the actions taken by individuals with particular identities both play strong roles in how we label other people (Heise, 1987). How semantic similarity and semantic association play into these more theoretically driven and advanced theories of identity labeling remains to be seen.

Acknowledgments

We would like to thank the MTurk workers for their efforts on this project, in particular Heather Wilcox, who gave some really insightful post-survey thoughts that improved the analysis of this data and helped to hone inquiries into future work. We also thank the reviewers for this paper, who offered several helpful critiques that greatly improved the article. We would also like to thank Lynn Smith-Lovin for her assistance in developing the surveys presented. Support was provided, in part, by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning and the ONR through a MINERVA N000141310835 on State Stability.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Camiel J. Beukeboom and others. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, 31:313–330.

Peter M. Blau. 1977. A macrosociological theory of social structure. *American journal of sociology*, pages 26–54.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

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. *arXiv preprint arXiv:1607.06520*.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Peter J. Burke. 1980. The self: Measurement requirements from an interactionist perspective. *Social psychology quarterly*, pages 18–29.

Mina Cikara and Jay J. Van Bavel. 2014. The Neuroscience of Intergroup Relations An Integrative Review. *Perspectives on Psychological Science*, 9(3):245–274.

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Jonathan B. Freeman and Nalini Ambady. 2011. A dynamic interactive theory of person construal. *Psychological review*, 118(2):247.

Trevor J. Hastie and Robert J. Tibshirani. 1990. *Generalized additive models*, volume 43. CRC Press.

David R. Heise. 1987. Affect control theory: Concepts and model. *The Journal of Mathematical Sociology*, 13(1-2):1–33, December.

David R. Heise. 2007. *Expressive Order*. Springer.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Kenneth Joseph, Wei Wei, and Kathleen M. Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from Twitter. In *2017 ACM Conference on Computer Supported Cooperative Work.(CSCW)*.

Kenneth Joseph. 2016. *New methods for large-scale analyses of social identities and stereotypes*. Ph.D. thesis, Carnegie Mellon University.

Sonia K. Kang and Galen V. Bodenhausen. 2015. Multiple Identities in Social Perception and Interaction: Challenges and Opportunities. *Annual Review of Psychology*, 66(1):547–574.

Ziva Kunda and Paul Thagard. 1996. Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103(2):284–308.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Timothy J. Owens, Dawn T. Robinson, and Lynn Smith-Lovin. 2010. Three faces of identity. *Sociology*, 36(1):477.

Andrew M. Penner and Aliya Saperstein. 2013. Engendering Racial Perceptions An Intersectional Analysis of How Social Status Shapes Race. *Gender & Society*, 27(3):319–344, June.

R Core Team, 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Marta Recasens, Eduard Hovy, and M. Antnia Mart. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152, May.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130.

Magnus Sahlgren. 2006. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Cosma Rohilla Shalizi. 2013. Advanced data analysis from an elementary point of view. *URL: http://www. stat. cmu. edu/cshalizi/ADAfaEPoV/13*, 24.

Lynn Smith-Lovin. 2007. The Strength of Weak Identities: Social Structural Sources of Self, Situation and Emotional Experience. *Social Psychology Quarterly*, 70(2):106–124, June.

Sheldon Stryker and Peter J. Burke. 2000. The past, present, and future of an identity theory. *Social psychology quarterly*, pages 284–297.

Amos Tversky and Itamar Gati. 1978. Studies of similarity. *Cognition and categorization*, 1(1978):79–98.

Emiel van Miltenburg. 2016. Stereotyping and Bias in the Flickr30k Dataset. In *Workshop on Computer Vision and Language Processing*.

Simon Wood. 2006. *Generalized additive models: an introduction with R*. CRC press.

Ziqi Zhang, Anna Lisa Gentile, and Fabio Ciravenga. 2013. Recent advances in methods of lexical semantic relatedness - a survey. *Natural Language Engineering*, 19(4):411–479, October.