Semi-supervised methods for expanding psycholinguistics norms by integrating distributional similarity with the structure of WordNet

Michael Mohler, Marc Tomlinson, David Bracewell, Bryan Rink
Language Computer Corp.,
Richardson, TX, USA
<michael,marc,david,bryan>@languagecomputer.com

Abstract

In this work, we present two complementary methods for the expansion of psycholinguistics norms. The first method is a random-traversal spreading activation approach which transfers existing norms onto semantically related terms using notions of synonymy, hypernymy, and pertainymy to approach full coverage of the English language. The second method makes use of recent advances in distributional similarity representation to transfer existing norms to their closest neighbors in a high-dimensional vector space. These two methods (along with a naive hybrid approach combining the two) have been shown to significantly outperform a state-of-the-art resource expansion system at our pilot task of imageability expansion. We have evaluated these systems in a cross-validation experiment using 8,188 norms found in existing pscholinguistics literature. We have also validated the quality of these combined norms by performing a small study using Amazon Mechanical Turk (AMT).

Keywords: Resource Expansion, Psycholinguistics, Distributional Similarity

1. Introduction

Psycholinguistics is the study of the acquisition, comprehension, and production of language by humans. Recently, the use of psycholinguistics norms – that is, a collection of scores which represent a typical human subject's reaction to or perception of a set of words from a variety of psycholinguistic perspectives – has come into prominent use with applictions in natural language processing (NLP) subfields as disperate as text simplification, personality modeling, and metaphor processing. For instance, Mairesse et al. (2007) showed that the use of vivid language (words with high concreteness and high imageability ratings) was a strong indicator of an extroverted personality. Likewise, age-of-acquisition, imageability, concreteness, meaningfulness, and familiarity norms have been used directly in the development of automated text generation systems (e.g. summarization systems) geared towards limited vocabulary groups such as children and other language learners (Kandula et al., ; Coster and Kauchak, 2011; Crossley et al., 2012). Further afield, research has been conducted studying the effect of concreteness on such tasks as semantic similarity/association (Hill et al., 2013), information retrieval (Tanaka et al., 2013), and word sense disambiguation (Kwong, 2008). Finally, imageability and concreteness have been shown to be prominent indicators of metaphoricity in the field of metaphor detection (Turney et al., 2011; Broadwell et al., 2013; Bracewell et al., 2014).

Unfortunately, research that makes use of such psycholinguistic norms remains dependent upon several small datasets produced sporatically over the last several decades which, taken together, represent only a fraction of the words present in the English language. This serves to significantly limit their usability when addressing problems in general-purpose NLP. In order to overcome this limitation, we have developed a two-part, semi-supervised methodology to expand existing psycholinguistics data for use by NLP researchers that approaches full coverage for the English language while maintaining high agreement with human assessments. This work has been carried out in support of our development of a multi-faceted metaphor detection system (Bracewell et al., 2014) which makes use of our expanded norms for word imageability, concreteness, arousal, dominance, and emotional valence. We here report only our work in the expansion of imageability norms, but the expansion process used for each attribute is identical.

The remainder of this work is organized as follows. In Section 2. we situate our contributions within the existing literature of psycholinguistic norm estimation. Then, in Section 3., we describe the dataset being used to evaluate our methods as well as the results of a mechanical turk study, validating and expanding the data. In Section 4. we describe our spreading activation and distributional similarity approaches to the problem. Then, we present our experiments in Section 5. and report results. Finally, in Section 6. we analyze the results of our experiments and propose future work.

2. Related Work

Existing research on the expansion of psycholinguistics resources can broadly be divided into two categories: those that make use of existing hierarchical structures of words (esp. WordNet) and those that use some measure of semantic similarity to predict the psycholinguistic characteristics of an unknown word. Most simply, Xing et al. (2010) estimated term concreteness for image retrieval by partitioning the WordNet hierarchy into abstract terms and concrete terms based solely on whether or not a concept was subsumed under the synset associated with "physical entity". Changizi (2008), in rating the organization of various lexical hierarchies, made use of the idea of a word's hypernymy level in the hierarchy as a rough estimate of its concreteness, while Sanchez et al. (2011) estimated concept generality (i.e., its abstractness) by counting the number of leaves subsumed by that concept in WordNet.

Feng et al., (2011) have attempted to predict the concreteness of unseen nouns using a supervised regression model. Using a total of 39 features including ontology depth, WordNet lexicographer files, sense counts, word frequency, and individual LSA dimension values, they were

able to achieve a correlation of 0.64 with human concreteness scores for nouns, but did not evaluate their methodology for other parts-of-speech. The approach of Broadwell et al., (2013) (which is similar to our spreading activation methodology) focuses on conservatively propagating known imageability ratings across a restricted set of WordNet semantic relations. Because of this, their methodology remains limited in coverage reaching only around 33% of the synsets in WordNet.

Among those approaches that make use of word similarity to estimate psycholinguistics norms, the approach of Turney et al., (2011) stands out. They iteratively and greedily select a total of 40 abstract and concrete prototypical terms from existing psycholinguistics literature. From these prototypes, they produce a score for an unseen word by combining the concreteness scores of those prototypes weighted according to their LSA similarity scores. Likewise, the $DIC - LSA$ system of Bestgen and Vincze (2012) propagates lexical norms for valence, dominance, arousal, imageability, and concreteness by taking the average score of the term's k-nearest neighbors in a reduced dimensionality vector space. Finally, Brooke and Hirst (2013) use several flavors of term similarity (including PMI, LSA, and LDA) to propagate a variety of stylistic features (including abstractness and concreteness) between unknown terms and a set of manually defined seeds.

Our methodology seeks to incorporate the observations uncovered within both strains of research. By integrating our methodology with the WordNet hierarchy, we are able to exploit the human knowledge encoded within it while providing full coverage of all English word senses for a variety of psycholinguistics norms. Then, by combining our WordNet-based approach with a model of distributional similarity, we are able to provide psycholinguistics ratings for out-of-lexicon words. This is especially crucial when working with languages for which well-developed lexicon resources are unavailable.

3. Data

The imageability of a word is defined as a measure of the extent to which that word brings to mind a sensory experience (i.e., something visual, auditory, tactile, etc.). In order to faciliate the evaluation of our techniques for dispersing imageability ratings, we have combined several existing imageability norms from a variety of sources. To begin with, we have taken imageability norms from the MRC Psycholinguistics Database (Coltheart, 1981) which represents standard foundational data being used in a variety of contemporary psycholinguistics research. Furthermore, we have supplemented this database with norms published in more recent research (Clark and Paivio, 2004; Cortese and Fugett, 2004; Schock et al., 2012; Friendly et al., 1982) so as to utilize high-quality human responses to the maximum extent possible.

In addition, we have carried out a pilot study using Amazon Mechanical Turk[1] with the explicit goal of providing norms for terms (one or more words) that are widely known, but are less common than those under focus in existing psycholinguistics research. In producing this supplemental dataset, we have attempted as much as possible to reproduce the experimental conditions as described in the Paivio (1968) study, one of the original studies included in the MRC Psycholinguistics Database. Participants were provided with a prompt introducing the concept of imageability and describing the task. The prompt was taken directly from the Paivio study, with a slight modification to the method of participant input (i.e. selecting instead of circling). Participants were also provided with an example consisting of a small set of words along with their approximate scores taken from the original study. They were then shown a set of words (with no context or part-of-speech information) and were asked to rate the words for imageability on a scale of 1 to 7.

Altogether, we produced human-quality imageability scores for 300 words or multi-word expressions. Of these, 100 terms shared a synset with one or more words included in the original MRC dataset, 100 terms could be linked to a synset of the MRC dataset using a single semantic relation (i.e. a sister term, hypernym, or hyponym relation), and 100 terms were more distantly related. These words were sampled randomly across all of WordNet within each imageability quartile (as determined by a baseline version of our spreading activation component) such that 25 terms from each set were taken from each quartile. Of these 25, we attempted to provide a variety of parts-of-speech with a target distribution of 12 nouns, 8 verbs, 4 adjectives, and 1 adverb in each group. The results of our study validate the use of Mechanical Turk for this purpose.

Pearson correlation between the results of our study and those of the original MRC dataset stood at 0.71 with a root mean squared error (RMSE) of 1.11 on a scale of 1 to 7. Two words stood out as representing a large discrepancy between our data and the MRC data: initiatory (5.54 in MRC and 1.40 in our study) and bewilder (5.10 in MRC and 2.00 in our study). A deeper analysis of these discrepancies showed that these words, which were not directly measured in the MRC, shared a synset with one or more words in the MRC which have a much more common, imageable sense i.e., maiden with initiatory and puzzle with bewilder.

This result led us to re-evaluate our original assumptions regarding the transfer of the gold-standard imageability scores onto WordNet synsets, and so we now address this issue by associating each norm with only a single WordNet synset chosen according to a simple heuristic. Specifically, we map the norm of a word to that word's most frequent sense in a predefined part-of-speech order. This means that we first attempt to link a score to a noun sense, if one is present. If there are none, we attempt to link the score to a verb sense, then to an adjective, and finally to an adverbial sense. Ratings mapped to senses within the same synset are then averaged and applied to the synset as a whole. We believe that this is a strong, context-free heuristic to associate a word (which has an associated norm) with a sysnet (which represents the physical or abstract thing actually being imaged by the study participants).

[1] www.mturk.com

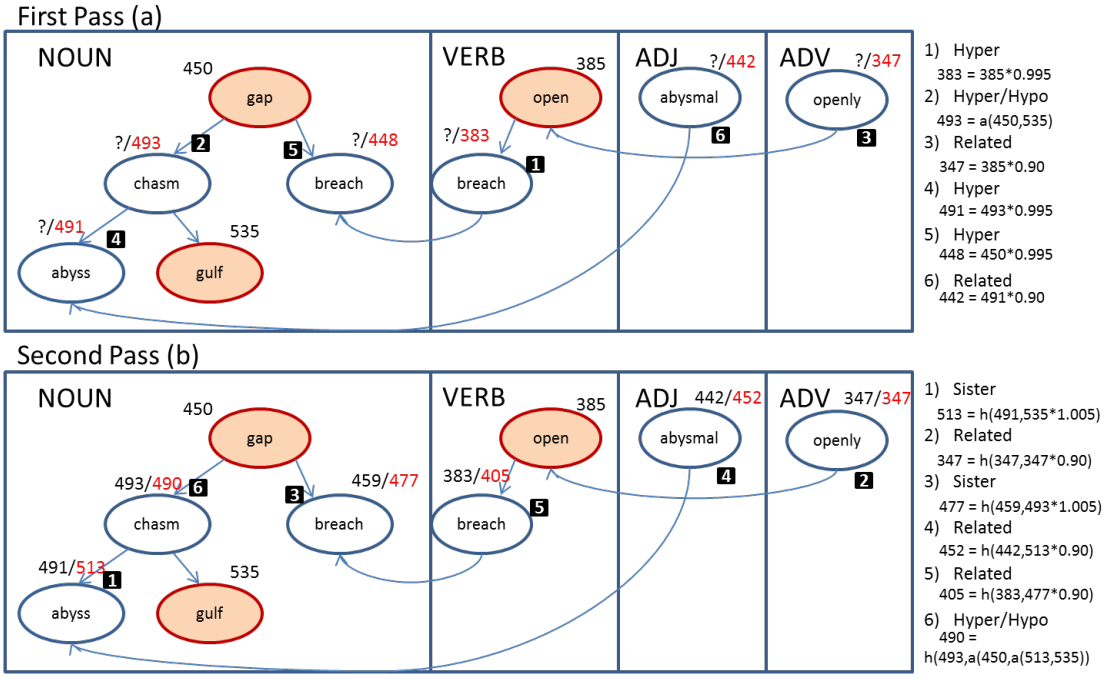

Figure 1: A visual representation of our spreading activation method for psycholinguistic rating expansion where shaded nodes represent WordNet synsets with gold-standard psycholinguistics norm, unshaded nodes represent all other synsets, and edges represent hypernymy/hyponymy relations (within a part-of-speech) or pertainym/derived term relations (across parts-of-speech). We show here two iterations of the algorithm. In each iteration, the synsets are randomly ordered (indicated by the number in the black boxes above). During the first iteration, the synsets are visited in the black-box order and updated by selecting one of the group types shown in the equations on the right-hand side (e.g. a 'hypernym' group for 'breach', then the 'hypernym/hyponym' group for 'chasm' and so on. The scores are updated according to the equations described in Section 4.1.. Once the first iteration is complete, each synset has received a score. On the second iteration, the process is repeated with a new random black-box order and new random group selections, which result in each synset receiving an updated score.

4. Methodology

Our current work advances the state-of-the-art in two ways using two independent components. The first component maps known imageability ratings onto WordNet synsets and, using a small set of semantic relations, disperses those ratings throughout the ontology using an iterative spreading activation process. The foundational observation motivation our development of the second component is that words with similar psycholinguistic properties occur in similar contexts, and so it is possible to predict the imageability of a word by considering only the norms of the word's nearest neighbors within a distributional vector space. Importantly, this component does not require a full ontological structure, thereby enabling our system to generalize to out-of-lexicon words and to perform effectively for languages with less well-developed language resources.

4.1. Spreading Activation

Beginning with a set of gold-standard imageability norms taken from the psycholinguistics literature (cf. Section 3.), we first assign ratings for each of these words onto a single WordNet synset as described above. These gold-standard ratings will be locked in place and not updated during the rest of our process. We then perform an iterative dispersal process that operates across five types of semantic relations – sister terms, hypernyms, hyponyms, pertainyms, and derived terms.

Our motivation for selecting these relations rests on three assumptions. First, it has been widely acknowledged in recent research (Changizi, 2008; Feng et al., 2011) that the generality/specificity scale encoded in a concept hierarchy (such as WordNet) is strongly correlated with measures of concept concreteness (and more loosely with imageability). Indeed, as a general trend in the MRC, it is clear that terms in the upper regions of the hierarchy (e.g. ANIMAL [575], PRODUCT [435]) are less imageable than their more specific descendants (e.g. COW [632], DOG [636], BOOK [591], MOVIE [571]). From this observation, we predict that the imageability of a word's hyponyms are a function of the word's imageability yielding a gradual increase as we descend the hierarchy. The the converse is true for its hypernyms.

Our second assumption is that the process of specification represents a roughly equivalent imageability transformation from a hypernym to each of its hyponyms, and so a term's rating can be estimated as that of a sister term in the hierarchy. Our third assumption is that words formed from existing words with a different part-of-speech (e.g. deverbal nouns, denominal adjetives) will inherit the majority of its imageability from its origin word. For instance, the verb "to dog <someone>" is vivid language insofar as it brings to mind the original noun "dog". We therefore predict that such terms will be slightly less imageable than their origin words as a general rule. From these assumptions, we define a multi-stage iterative process to disperse imageability rat-

ings across all of WordNet. Two iterations of this process are shown in Figure 1.

During each iteration x of the process, the imageability rating associated with every synset in WordNet is updated in a random order.[2] For each synset, S_i, that does not contain a gold-standard norm, one of the following three groups of synsets is randomly selected: those that are siblings of S_i, those that are hypernyms or hyponyms of S_i, and those that are pertainyms of S_i (or derivationally-related terms). The ratings associated with this group, G, are then used to assign an imageability score, S_i^x, for the synset as follows:

$$S_i^{x+1} = \begin{cases} f(G) & \text{if x=0} \\ h(S_i^x, f(S,G)) & \text{if x>0} \end{cases}$$

$$f(G) = \begin{cases} k_G * ave(p(G)) & \text{if G!=ANS+DES} \\ ave(p(ANS), p(DES)) & \text{otherwise} \end{cases}$$

$$p(G) = ave(syns(G))$$

where $h(x, y)$ represents the harmonic mean of x and y, $ave(x, y)$ represents the arithmetic mean of x and y, $syns(G)$ represents the current ratings for all synsets within group G, and k_G is a scaling factor based on the type of the group, G. [3]. In order to maintain a bound on the scores, values are clipped to the range [1 to 7] after each update. After a fixed number of iterations, the result of this process is a total covering of WordNet with imageability scores applied at the synset level.

4.2. Distributional Similarity Expansion

As an alternative method that does not depend on an existing, well-defined lexical ontology, we have developed a vector similarity approach which is based on the assumption that words which occur in similar contexts will share a variety of psycholinguistic features including imageability. For the purposes of this work, we define context in a manner similar to that of Lin (1998), who defined a word vector space based on relations and a word's cooccurrents via each relation. In particular, a word's vector was defined by the number of times that word occurred within a set of (word, relation, word') tuples. Our vector space (which we refer to as DepVec space) extends Lin's space by incorporating information about relational (e.g. selectional) preference as measured by the G-test score. This score is a measure of the associativity of two things defined as:

$$G = 2 \sum_i O_i \cdot \ln(O_i / E_i)$$

where O_i is the observed frequency in a cell of the 2x2 cooccurrence matrix and E_i is the expected frequency in that cell according to the null hypothesis (i.e. that the two words are independent across the given relation). This cooccurrence matrix was computed as a result of a dependency parse of the 13 million English documents in the ClueWeb09 corpus.[4] A sample of the dominant relations (i.e. DepVec dimensions) associated with the word "cure" is shown in Figure 2.

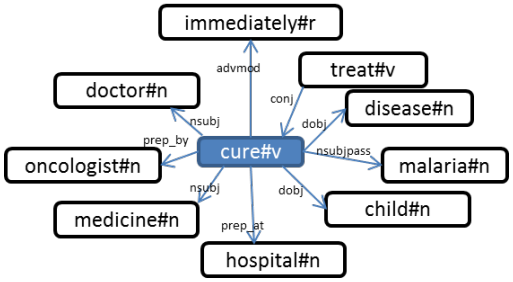

Figure 2: A graphical representation of the dominant dependency relations for the word "cure".

In order to assess the imageability of an unknown word, x, we first generate a list of the 1,000 words that are most distributionally similar to the target word (i.e. its nearest neighbors in the DepVec space). Next, we take the intersection of that list with the set of words from our gold-standard psycholinguistics data. From this intersection, $SIM(x)$, we then calculate, $I(x)$, the imageability of our x based on a rank-weighted average of the imageability of all words, $w_i^x \in SIM(x)$, as follows:

$$I(x) = \frac{\sum_{w_i^x \in SIM(x)} \frac{sim(x, w_i^x)}{rank(w_i^x)}}{\sum_{w_i^x \in SIM(x)} \frac{1}{rank(w_i^x)}}$$

where $rank(w_i^x)$ represents the 1-based rank of w_i^x in $SIM(x)$.

In addition, we have experimented with an alternative approach which weights the imageability scores based on the similarity (not the rank) between the known words and our target word. In particular, we computed the weighted average of similar words (of known imageability) such that the weight was calcuated as the cosine similarity between the vectors of the target word and a known word. Preliminary results showed that the rank-weighted similarity approach performed slightly better than the similarity-weighted approach for English and so we report only the rank-weighted methodology.

4.3. Hybrid Expansion

In the hopes of combining the full WordNet coverage of our spreading activation component with the high-quality and the out-of-lexicon capability of the distributional similarity component, we have combined the two into an integrated hybrid system. In particular, we attempt to compute a score for an unknown word using both methods. If both methods are able to provide a score[5], they are scaled and combined

[2] We hypothesize that by employing a randomized expansion, we avoid the introduction of bias in selecting an initial norm to drive the expansion, and that by iterating this process, the accumulated effects of the random aspects will result in convergence near the true rating for each synset.

[3] Initially, we expected that scores would be slightly increased for hyponyms (DES), slightly lower for hypernms (ANS) and pertainyms or derivationally related words (REL), and roughly equal for sister terms (SIS). However, based upon a parameter-tuning grid search over three iterations, we have selected the following for k: ($k_{SIS} = 1.005, k_{ANS} = 0.995, k_{DES} = 0.900, k_{REL} = 0.900$). We use these values for all experiments in Section 5.

[4] http://lemurproject.org/clueweb09/

[5] In particular, the spreading activation method will miss any words not found in WordNet, and the distributional similarity

using a scaling factor determined based on the experiments described in Section 5.1..

5. Experimental Results

We here describe three experiments to analyze the quality of our methodologies. In the first, we show the effects of adjusting two high-level parameters of our system – namely the number of iterations to perform in the WordNet-based expansion component and the relative weights of the two components in the full hybrid system. In the second experiment, we compare our methodology against several baselines, heuristic methods, and an existing state-of-the-art system in a cross-validation experiment over the supplemented MRC imageability data. Finally, we compare against the same systems using the psycholinguistics data as training while evaluating on the dataset described in Section 3.. In each experiment, we report performance using Pearson's correlation coefficient (r) and the root mean square error (RMSE).

5.1. Parameter Tuning

Figure 3 shows the results of varying the number of iterations for our spreading activation methodology from 1 to 10. The trendlines indicate that the overall quality of the ratings (as measured by the Pearson correlation coefficient[6]) improve with more iterations for adjectives and especially for adverbs, while additional iterations seem to deteriorate the quality of the ratings slightly for both nouns and verbs. Based on these results, we have chosen to limit the number of iterations to three for the remainder of the experiments in order to minimize the negative effects on nouns and verbs while allowing for improvement for the other parts-of-speech.

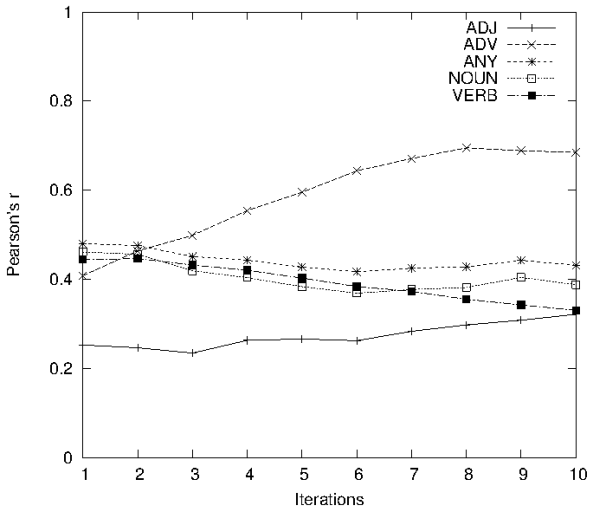

Figure 3: Correlation in the cross-validation experiment for various parts-of-speech as the number of iterations for the WordNet-based expansion methodology are varied.

method will miss words that cannot be found in our parsed relations corpus or those that have no similar words in the psycholinguistics data.

[6]The RMSE scores are omitted here for the sake of readability, but the trends are the same for both measures.

In Figure 4, we see the effects of varying the scaling factor in the combination of our spreading activation method and our distributional similarity method. Weights ranged from 0.0 (all spreading activation) to 1.0 (all distributional similarity). It is clear from this experiment that the optimal weight is at neither extreme, which suggests that the two methods, used in conjunction, serve to complement one another and to bring out the strengths of each. Based on the results of this experiment, we have selected a weight of 0.85 for use in the remaining experiments.

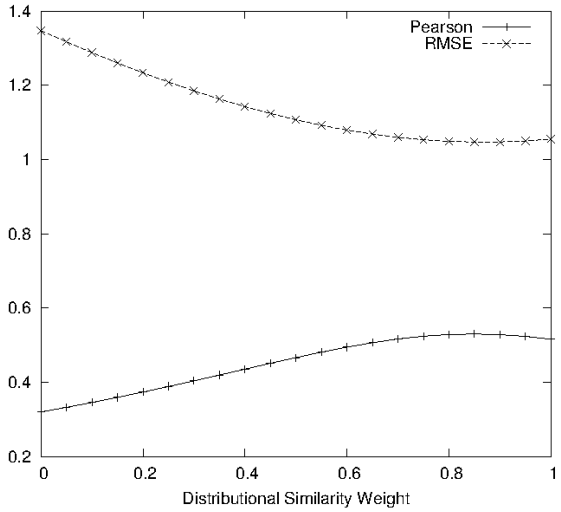

Figure 4: Correlation and RMSE in the cross-validation experiment when varying the relative weight of WordNet-based and distributional similarity methods.

5.2. Cross-Validation Experiment

In our next experiment, we perform a 5-fold cross-validation over our supplemented MRC norms from Section 3. using a variety of expansion methods. These methods include our distributional similarity method (DISTSIM), our spreading activation method (SPREAD), and our hybrid system (HYBRID). In addition, we compare our systems to other work by reimplementing the methodology proposed by Broadwell et al.(2013). Furthermore, we have implemented two heuristic methods inspired by previous work. One (AVE_LEXFILES) takes the average rating from the norms in the training set for all synsets with the same WordNet lexicographer file, while the other (AVE_DEPTH) takes the average rating for all synsets at the same depth in the ontology. Finally, we compare against three naïve baselines: one that produces a random score within the 1 to 7 range (RANDOM), one that randomly samples the gold-standard data (RANDOM_NORM), and one that assigns the average score of the gold-standard data to all unseen synsets (AVE_NORM).

The results of this experiment (producing imageability ratings for a total of 8,188 words) are shown in Table 1 with the methods categorized as partial-coverage methods (DISTSIM and Broadwell et al.(2013))[7], full-coverage methods (HYBRID, AVE_LEXFILES,

[7]For the partial-coverage methodologies, non-covered words are ignored with no penalty

SPREAD, AVE_DEPTH), and baselines (AVE_NORM, RANDOM_NORM, RANDOM).

Taken as a whole, the distributional similarity approach (DISTSIM) outperformed all other approaches with our full-coverage hybrid methdology slightly behind. At the same time, the spreading activation approach resulted in higher quality scores for adjectives and adverbs, suggesting that pertainym and derivational relations represent a more fitting conduit for imageability transfer than context for these parts-of-speech. One suprise from these experiments is the relatively high quality of the lexicographer file heuristic (AVE_LEXFILES) which has full coverage and results in significantly less error both the Broadwell et al.(2013) method and the (SPREAD) method – overall and for nouns in particular.

5.3. Held-out Experiment

In our final experiment, we used the full supplemented MRC norms as gold-standard data and evaluated using the Mechanical Turk data described in Section 3. as a held-out set. Recall that this data was produced in order to provide norms for words that are uncommon, but well-known (e.g. "bulletproof vest", or "attractiveness") which differs qualitatively from the majority of existing norms. Altogether, this set contains 164 words not found in the supplemented MRC norms. The results of this experiment are shown in Figure 2. This set had too few adverbs (5) to evaluate separately.

These results (on a different class of words than those in the experiment of Section 5.2.) show a mixed result in the comparison of our two systems and Broadwell et al.(2013). As we can see, the Broadwell et al.(2013) system (ignoring the coverage differences) appears to outperform both of our systems for nouns and verbs when performance is measured by correlation, while the opposite result is true when measuring error. This highlights the need for multi-faceted evaluation as either measurement alone fails to sufficiently describe the quality of the results. It is perhaps unsuprising, as the more conservative Broadwell et al.(2013) method tends to propagate ratings across only the highest-confidence relations making no attempt to achieve a high coverage.

6. Conclusions

In this work, we have introduced a two-part methodology for estimating psycholinguistics ratings such as word imageability for words not found in existing psycholinguistics norms. Our first method, based on a spreading activation dispersal of existing norms throughout WordNet approaches full coverage of the English language. Our second method, which uses the most distributionally similar words among those with known ratings to predict the rating for an unknown word, has been shown to significantly outperform a suite of baseline methodologies, informed heuristics, and an existing state-of-the-art system. Working in tandem, we have shown that these two methods combine high coverage with high quality for use in general-purpose NLP.

As a next step, we intend to incorporate several of the observations uncovered over the course of this work. First, we intend to make explicit use of the lexicographer files and the synset depth information (e.g. as part of a regression model) in our WordNet-based expansion component as this information has been shown to correlate positively with human imageability ratings. Second, we will refine the expansion process to estimate scores for words with particular attention paid to part-of-speech, as our results indicate that the quality of the ratings are uneven across different parts-of-speech and that some methods may be more appropriate than others for a given part-of-speech. We further intend to evaluate our methodologies in a cross-lingual context so as to evaluate the effectiveness of transferring psycholinguistics information across languages (i.e. via a linked synset) thereby making available the psycholinguistics data for robust multilingual natural language processing.

Acknowledgments

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0025. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. References

Bestgen, Y. and Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior research methods*, 44(4):998–1006.

Bracewell, D., Tomlinson, M., Mohler, M., and Rink, B. (2014). A tiered approach to the recognition of metaphor. In *Computational Linguistics and Intelligent Text Processing*.

Broadwell, G. A., Boz, U., Cases, I., Strzalkowski, T., Feldman, L., Taylor, S., Shaikh, S., Liu, T., Cho, K., and Webb, N. (2013). Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 102–110. Springer.

Brooke, J. and Hirst, G. (2013). Hybrid models for lexical acquisition of correlated styles.

Changizi, M. A. (2008). Economically organized hierarchies in wordnet and the oxford english dictionary. *Cognitive Systems Research*, 9(3):214–228.

Clark, J. M. and Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):371–383.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Cortese, M. J. and Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, 36(3):384–387.

Coster, W. and Kauchak, D. (2011). Learning to simplify sentences using Wikipedia. In *Proceedings of the Work-*

Method	coverage	Any POS (8,188) r	Any POS (8,188) RMSE	Noun (3,111) r	Noun (3,111) RMSE	Verb (3,133) r	Verb (3,133) RMSE	Adj (1,593) r	Adj (1,593) RMSE	Adv (351) r	Adv (351) RMSE
DISTSIM	94%	**0.56**	**1.009**	**0.59**	**1.014**	**0.53**	**1.024**	0.43	0.970	0.49	1.048
Broadwell et al.(2013)	37%	0.34	1.409	0.34	1.388	0.22	1.453	0.20	1.435	**0.61**	1.148
HYBRID	100%	0.53	1.047	0.52	1.094	0.51	1.051	**0.45**	**0.949**	0.51	**1.018**
AVE_LEXFILES	100%	0.34	1.177	0.41	1.164	0.21	1.237	0.10	1.104	0.33	1.063
SPREAD	100%	0.32	1.347	0.34	1.322	0.19	1.400	0.26	1.325	0.26	1.163
AVE_DEPTH	100%	0.22	1.228	0.20	1.260	0.12	1.263	0.08	1.112	0.30	1.134
AVE_NORM	100%	0.11	1.264	0.10	1.304	0.09	1.275	0.08	1.147	0.27	1.316
RANDOM_NORM	100%	0.04	1.717	0.04	1.762	0.04	1.706	0.01	1.660	0.15	1.670
RANDOM	100%	0.03	2.090	0.03	2.189	0.03	2.060	-0.01	1.987	0.16	1.899

Table 1: Comparison of various methods and baselines in 5-fold cross validation of existing psycholinguistics norms (8,188 words total).

Method	coverage	Any POS (164) r	Any POS (164) RMSE	Noun (98) r	Noun (98) RMSE	Verb (44) r	Verb (44) RMSE	Adj (17) r	Adj (17) RMSE
DISTSIM	37%	0.33	**1.265**	0.37	1.511	0.20	1.179	**0.51**	**1.043**
Broadwell et al.(2013)	42%	**0.54**	1.360	**0.58**	1.329	**0.52**	1.399	N/A	N/A
HYBRID	100%	0.48	1.322	0.45	1.451	0.29	**1.084**	0.45	1.250
AVE_LEXFILES	100%	0.51	1.340	0.53	1.398	0.32	1.238	0.28	1.431
SPREAD	100%	0.32	1.347	0.34	**1.322**	0.19	1.400	0.26	1.325
AVE_DEPTH	100%	0.43	1.428	0.41	1.536	-0.07	1.315	0.40	1.211
AVE_NORM	100%	N/A	1.498	N/A	1.620	N/A	1.315	N/A	1.211
RANDOM_NORM	100%	-0.01	1.947	-0.06	2.090	-0.01	1.611	-0.09	2.000
RANDOM	100%	0.10	2.126	0.12	2.166	-0.13	2.083	0.09	2.007

Table 2: Comparison of various methods and baselines on held out data from Amazon Mechanical Turk (164 words total).

shop on Monolingual Text-To-Text Generation, pages 1–9. Association for Computational Linguistics.

Crossley, S. A., Allen, D., and McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1):89–108.

Feng, S., Cai, Z., Crossley, S. A., and McNamara, D. S. (2011). Simulating human ratings on word concreteness. In *FLAIRS Conference*.

Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14(4):375–399.

Hill, F., Kiela, D., and Korhonen, A. (2013). Concreteness and corpora: A theoretical and practical analysis. *CMCL 2013*, page 75.

Kandula, S., Curtis, D., and Zeng-Treitler, Q.). A semantic and syntactic text simplification tool for health content.

Kwong, O. Y. (2008). A preliminary study on the impact of lexical concreteness on word sense disambiguation. In *PACLIC*, pages 235–244.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500.

Paivio, A., Yuille, J. C., and Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(12):1.

Sánchez, D., Batet, M., and Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297–303.

Schock, J., Cortese, M. J., and Khanna, M. M. (2012). Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods*, 44(2):374–379.

Tanaka, S., Jatowt, A., Kato, M. P., and Tanaka, K. (2013). Estimating content concreteness for finding comprehensible documents. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 475–484. ACM.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.

Xing, X., Zhang, Y., and Han, M. (2010). Query difficulty prediction for contextual image retrieval. In *Advances in Information Retrieval*, pages 581–585. Springer.