

Modeling Word Meaning: Distributional Semantics and the Corpus Quality-Quantity Trade-Off

*Seshadri Sridharan*¹ *Brian Murphy*²

(1) Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

(2) Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA

seshadrs@cs.cmu.edu, brianmurphy@cmu.edu

ABSTRACT

Dictionaries constructed using distributional models of lexical semantics have a wide range of applications in NLP and in the modeling of linguistic cognition. However when constructing such a model, we are faced with range of corpora to choose from. Often there is a choice between small carefully constructed corpora of well-edited text, and very large heterogeneous collections harvested automatically from the web. There may also be differences in the distribution of genres and registers in such corpora. In this paper we examine these trade-offs by constructing a simple SVD-reduced word-collocate model, using four English corpora: the Google Web 5-gram collection, the Google Book 5-gram collection, the English Wikipedia, and collection of short social messages harvested from Twitter. Since these models need to encode semantics in a way that approximates the mental lexicon, we evaluate the felicity of the resulting semantic representations using a set of behavioral and neural-activity benchmarks that depend on word-similarity. We find that the quality of the input text has a very strong effect on the performance of the output model, and that a corpus of high quality at a small size can outperform a corpus of poor quality that is many orders of magnitude larger. We also explore the semantic closeness of the models using their mutual information overlap to interpret the similarity of corpus texts.

KEYWORDS : VECTOR SPACE MODELS, DISTRIBUTIONAL SEMANTICS, CORPUS SIZE, CORPUS GENRE, CORPUS QUALITY, NEUROSEMANTICS, WORD SIMILARITY

1 Introduction

Distributional semantic models (DSM) or distributional similarity models (Landauer, 1997) are unsupervised models based on the assertion that the meaning of a word can be inferred to some extent based on its distribution in the text. They are high dimensional vector space representations that encode the semantics of words learnt from a statistical analysis of the context they appear in. Word level dictionaries constructed using DSMs find use in many computational linguistics and cognitive science applications (Leacock, 1993; Bellegarda, 2000; Mitchell, 2008). To build these models in a given language, there is typically a choice among several source corpora. Often there is a choice between small well-curated text of good composition, and very large easy to collect text that is of inferior composition. In addition, there are choices along the dimensions of language style, genre and register. What kind of a corpus is most representative of a person's language experience? Is colloquial text more preferable than the formal variety? How does corpus size affect the model learnt? In this paper we attempt to identify the trade-off between source corpus size and quality, measured based on their performance in modeling the mental lexicon. Multiple behavioural and neurosemantic tests are used for this evaluation. As additional explorations, we study the effects of dimensionality on model performance, and the mutual similarity by word categories among models derived from various corpora.

There is ample literature analysing the effects of feature types, normalization, dimensionality, pruning, among other factors, on distributional semantics (Bullinaria and Levy 2007; Murphy, 2012). Quantitative and qualitative comparison of corpora based on the surface text has been performed as well (Kilgarriff 2001, 2012). But, to our knowledge, there is not any systematic analysis of the effect of the corpus quality on distributional semantics. Authors have expressed that it is not adequate to explore the effects of size on model quality, it is important to analyse the effects of corpus quality as well (Bullinaria and Levy 2007; Lindsey, 2007). Although a wide range of corpora have been used to build DSMs, variation in modeling parameters, processing techniques and evaluation metrics used by the authors makes a direct comparison of corpus quality unfeasible.

In this paper, we build simple SVD-reduced word-collocate models using four English corpora that differ considerably in quality, size and composition. We employ simple word co-occurrence based models rather than the more complex ones (such as dependency or document models) because it is possible to build word-collocate models for most languages and corpora that are available. More importantly, the goal of the paper is to arrive at general reliable performance trends to address the quality-quantity trade off, and not to obtain the very best performance possible. Thus, we employ more generic, commonly used methods and evaluation metrics in our experiments. We find that the quality of the input text has a very strong effect on the performance of the output model, and that a corpus of high quality at a small size can outperform a corpus of poor quality that is many orders of magnitude larger. And, we also explore the reasons for the relative performance of different corpora, in terms of the mutual similarity of the semantic spaces described by their corresponding models.

1.1 Characteristics of a Textual Corpus

Textual corpora vary along many dimensions. Microblogs are colloquial, abbreviated, have varied grammar, misspellings, and emoticons. Duplication of messages propagated by the social network (virality) is a phenomenon specific to this domain. On the other hand, books and news text are extremely formal, diligently edited content with superior use of the language. They are practically devoid of any spelling errors, adhere to conventional grammar and discuss a broad range of topics. Encyclopaedic sources contain factual accounts of entities in the world that go through the highest scrutiny by authors. They are well edited, and the use of language within an article is constrained to the subject of discussion, with pockets of rare terms within articles rather than a more even distribution across documents. Webpage characteristics generally are a mixture of all of the above. Content on the Internet is also skewed in its representation of topic and genre – for example computing topics may be over-represented.

The most desirable corpus to learn a cognitively plausible semantic model would be the one that is representative of the language experience of a native speaker. But, every corpus in some way is an idiosyncratic sample of the language, with biases of grammar, style and vocabulary, which may affect the semantic model that can be derived from it. For our experiments we consider four widely used research corpora that represent the major characteristics described above. They are available in many languages and in considerable sizes. We use Twitter messages (or tweets), Google Web n-grams, Google Books n-grams and Wikipedia articles. Tweets are short snippets of microblog text exchanged within a social network. The content tends to be biased towards the most trending news events and personal conversations. Webpages are online documents that are intended to be information resources. They are composed of heterogeneous data sources ranging

from product pages to blog posts to news articles etc. Books are works of literature that are carefully created by authors and typically edited by reviewers. The high quality text spans multitude of genres, topics and writing styles. The rest of the paper is structured as such: Section 2 describes the acquisition and pre-processing of corpora, process of building the semantic models and evaluating them. Section 3 details the experiments that vary the SVD dimensions and corpus sizes. Section 4 interprets those findings.

2 Methods

2.1 Collecting and Preprocessing Corpora

The Google Web corpus (Web) (Brants, 2006) contains n -grams of length up to 5 generated from publicly accessible Web pages. The Google Books dataset (Books) containing n -grams up to length 5 is extracted from a combination of dialects and genres, including American and British English, and both fiction and non-fiction. The Wikipedia corpus (Wiki) is a recent version, the July 2012 dump of the English encyclopaedia. Only running article text was extracted for use, with editing records, navigational text and other meta-data removed. The tweets corpus (Twitter) is a collection of 207 million public tweets collected from the twitter firehose over a 16-month period ranging from May 2009 to August 2010, a subset of the corpus collected by O'Connor et al., (2010). Among these, only tweets with five or more standard English words were retained, to discard non-standard utterances (e.g. telegraphic speech), and messages in other languages. The English word list used to filter the tweets contains the top 100K words in the American National Corpus (ANC). To avoid biases that reposting of messages may cause, duplicate posts were also discarded. All web links in the tweets were replaced with the token “[LINK]”, all usernames were replaced with “[PERSON]” and all hash-tags were stripped off the ‘#’ character and treated as normal tokens. After tokenizing the running text based on whitespaces, the tokens in all corpora were converted to lower case and only tokens composed solely of letters and internal punctuation were considered. No stemming or spelling correction was performed in the interest of impartiality towards all corpora.

	WIKIPEDIA	TWITTER	BOOKS	WEB
SIZE (ratios)	1 X	~1.2 X	~100 X	~200 X
LEXICAL DIVERSITY	483 k	736 k	135 k	206 k
CURATION	Very High, Peer Reviewed, Updated Frequently	None. High rate of typos and non standard language	Professionally edited.	Mix
REGISTER	Very Formal, Reporting Fashion	Very Informal, Colloquial	Formal, Narrative style	Mix
OBJECTIVITY	Completely Factual	More Opinions	More Fictional	More factual

TABLE 1 – Corpus facts and characteristics

After pre-processing, we found the Web, Books, Twitter and Wikipedia corpora to have 353.4 billion, 199.4 billion, 2.1 billion and 1.7 billion tokens respectively (see Table 1 for the same sizes expressed as ratios). Since the Google Books and Web corpora are available only as n -grams with a maximum sequence length of five, all other corpora were also reduced to five grams. Hence, all textual statistics were gathered from a fixed text window of 4 lower-case tokens either side of the target word of interest, which is in the mid-range of optimal values found by various authors (Lund and Burgess, 1996; Rapp, 2003; Sahlgren, 2006). Since the corpora are substantially different from each other, it is not feasible to use a common word-list as the vocabulary that would be equally suitable for all. As a result, we compiled a vocabulary specific to each corpus, taking all frequency-sorted tokens to achieve 99% token-coverage of that corpus (Table 1 shows the vocabulary size as lexical diversity).

2.2 Models of Semantics

All the models described here were subjected to a common pre-processing pipeline. Raw lower-case word co-occurrences were extracted in the ± 4 -word window. In the case of n -gram models that include a pre-applied frequency cut-off for rare tuples, a smoothing strategy was used to approximately reconstruct these missing counts. The 99% token-coverage vocabulary, and a subsequent 99% co-occurrence cut-off discarded low-frequency noisy counts before frequency normalization with PPMI (positive pointwise-mutual-information). The dimensionality of each word/collocate matrix was then reduced with singular value decomposition (SVD), taking the resulting left-singular vector as the vector-space representation for each word in the vocabulary.

$$\text{PPMI}_{wf} = \begin{cases} \text{PMI}_{wf} & \text{if } \text{PMI}_{wf} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\text{PMI}_{wf} = \log \left(\frac{p(w, f)}{p(w)p(f)} \right) \quad (2)$$

Positive Pointwise-mutual-information (1,2) is used as an association measure to normalize the observed word co-occurrence frequency $p(w,f)$ for the varying frequency of the target word $p(w)$ and its features $p(f)$. PPMI up-weights co-occurrences between rare words, yielding positive values for collocations that are more common than would be expected by chance, and discards negative values that represent patterns of co-occurrences that are rarer than one would expect by chance (i.e. if word distributions were independent). PPMI has been shown to perform well for a range of model types (Bullinaria and Levy, 2007; Turney and Pantel, 2010; Murphy, Talukdar and Mitchell, 2012). To filter out the noisy low frequency co-occurrences, we consider only those types that yield a 99% co-occurrence token coverage over all co-occurrence tokens. Filtering this Zipfian distribution also reduces the data to a manageable size.

The n -grams found in the Web and Books corpora were pre-filtered to different extents based on their counts. We observed that the lower the n -gram order, the lower the cut-off counts. So, to get a more accurate estimate of the co-occurrences in the original unfiltered corpus, we calculate the co-occurrence count using n -grams of all orders (two to five) rather than using only the five-grams. For a particular co-occurrence ab (where a and b are the co-occurring words), we calculate its scaled count c_{ab} (3) using NG_{ab} , the set of all n -grams that contain words a and b . $o(x)$ is the order of the n -gram x and $d(x,a,b)$ is the distance between words a and b in the n -gram. This scaled value is an approximation of the actual count a co-occurrence type would have when counted within five-grams from the original unfiltered corpus. This scaled count better

approximates the original count than the pre-filtered counts in the corpus five-grams. Preliminary evaluations on these corpora and the tests described later suggest that using these scaled counts to calculate the PPMI scores yields a better performance.

$$c_{ab} = \sum_{x \in NG_{ab}} \frac{5 - d(x, a, b)}{o(x) - d(x, a, b)} \quad (3)$$

Once the PPMI scores are obtained for all co-occurrences, every word has an associated vector containing the PPMI scores of that word with every word in the vocabulary. A singular value decomposition (SVD) is applied on the PPMI matrix to identify the k dimensions within each model with the greatest explanatory power, which also has the effect of combining similar dimensions (such as synonyms and inflectional variants) into common components, and discarding more noisy dimensions in the data. This gives us a vector of length k for every word in the vocabulary. We use these vector space models of word semantics (the word level dictionaries) produced by the SVD to perform behavioral tests and brain tests where we explore performance of models by varying the number of dimensions and the corpus size, one at a time.

2.3 Evaluating the Semantic Models

2.3.1 Neurolinguistic Decoding

Since neurosemantic tests require models to test directly on the brain activity associated with language, we believe they are a good approach to test models of the lexicon. The dataset used here is that reported in Mitchell et al., (2008) and released publicly as part of the First Workshop on Computational Neurolinguistics (Murphy et al., 2010). The functional MRI (fMRI) data had been recorded from 9 participants while they performed a property generation task. The stimuli were line-drawings, accompanied by their text label, of 60 everyday concrete concepts such as *ant*, *apartment*, *car*, *lettuce*, *hand*, *glass*. Each participant's data contained a time-course for each of approximately 20 thousand voxels (three-dimensional pixels, or neural data points), and multiple presentations of the same concept had been averaged to yield a single brain image for each concept. Following the analytical paradigm of (Mitchell et al., 2008), we use a linear model to predict the brain activity for a particular concept (4). For each participant and selected voxel, we train a model where the level of activation of the latter in response to different concepts is approximated by a regularized linear combination of their semantic features where f is the vector of activations of a specific fMRI feature for different concepts, the matrix C contains the values of the semantic features for the same concepts, β is the weight for each of those (corpus-derived) features, and λ tunes the degree of regularization.

$$f = \mathbf{C}\beta + \lambda\|\beta\|^2 \quad (4)$$

The linear model is estimated with a least squared errors method and $L2$ regularization, selecting λ over the range 0.0001 to 5000 using Generalized Cross-Validation (see Hastie et al., 2011, p.244). The activation of each fMRI voxel in response to a concept unseen during training is then predicted by the weighted sum of the values on each semantic dimension, building a picture of expected neural activity response for an arbitrary concept. We use the leave-2-out paradigm as used by Mitchell et al. (2008), in which a linear model for each neural feature is trained in turn on all concepts minus 2, having selected the 500 most stable voxels in the training set. For each of the 2 left-out concepts, we try to match the predicted and observed activations, using the cosine distance between the model-generated estimate of fMRI activity and that observed in the experiment. The score reported is the classification accuracy over the 1770 comparisons (60 select 2) by 9 participants.

2.3.2 Behavioral Measures

Since behavioral tests of language semantics capture human judgments based on their language experience, we believe they are a reasonable way to benchmark the different word dictionaries we generate. We apply commonly used behavioral tests of semantic knowledge (see e.g. Bullinaria and Levy 2007, Baroni et al 2010) to measure the quality of the corpus-derived models. Figure 1 depicts the distribution of the test vocabularies across the four corpora. All these tests involve pairwise comparison between two vectors, either corresponding to a pair of words, or between a word vector and a cluster centroid. We use the commonly used cosine geometric measure (Landauer and Dumais, 1997; Levy and Bullinara 2007) to calculate the distance between two vectors in the model’s vector space, independent of scaling. The distance measure is one minus the cosine of the angle between the two vectors t and e (5).

$$\text{cosine_similarity}(t, e) = 1 - \frac{\sum_{i=1}^n t_i e_i}{\sqrt{\sum_{i=1}^n (t_i)^2} \sqrt{\sum_{i=1}^n (e_i)^2}} \quad (5)$$

The TOEFL test, initially introduced by Landauer and Dumais (1997) consists of eighty multiple choice questions from the synonym portion of the TOEFL test. The test vocabulary consists of 203 adjectives, 96 abstract nouns and 82 verbs. The questions consist of a target word and four other word choices, including a synonym, and three distracters. (e.g. ‘Which of the following is closest in meaning to *prominent*: *battered*, *ancient*, *mysterious* or *conspicuous*?’). To evaluate a semantic model, we choose the word with the smallest cosine distance as the answer. The score reported is the answer-accuracy over all 80 questions.

The Rubenstein Goodenough (Rubenstein et al., 1965) and the WordSim (Finkelstein et al., 2002) datasets are comprised of word-pairs with corresponding a semantic-similarity score. R&G has 65 concrete noun pairs and WordSim has a mix of 203 nouns (concrete and abstract) and adjectives. The similarity scores for the word pairs (e.g. gem - jewel, 3.940) are values averaged over similarity judgments provided by multiple human judges. Modeled similarity scores for each word pair are generated using the cosine distance between the vectors in the semantic space. The test scores reported are Spearman correlation coefficients ρ between the similarity estimates x generated by the model and the gold standard similarity measures y (7).

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}. \quad (6)$$

The Battig (Battig and Montague, 1969) and AAMP (Almuhareb and Poesio, 2004) tests have pairs of a word and its immediate superordinate category (e.g. aeroplane – vehicle, anger – feeling). The Battig dataset is composed of 82 concrete words and AAMP has a mix of 402 concrete and abstract words. The CLUTO clustering toolkit (Karypis 2003) is used to cluster the word vectors using cosine distance and the toolkit’s default parameters to obtain as many clusters as there are word-categories in the test. The score reported is the overall cluster purity P , the sum of the purities of individual clusters (P_r) calculated (8). The purity of a cluster is the fraction of its members that belong to the most representative (i.e. plurality) category c .

$$P = \sum_{r=1}^k \frac{n_r}{n} P_r \quad (7) \quad P_r = \frac{1}{n_r} \max_c (n_r^c) \quad (8)$$

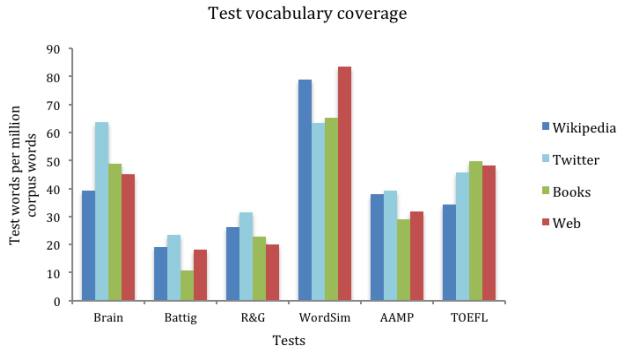


FIGURE 1 – Coverage per test-word as average word count per million corpus tokens (wpm)

3 Results

Our goal is to compare the performance of the models learnt from the four corpora in an impartial manner. Hence, we learn models on the corpora subsampled at the same size, at a particular dimensionality that is suitable to all corpora. To determine the appropriate number of SVD dimensions for the corpora, we compile test scores for the corpus models at the original sizes, varying the number of SVD dimensions (Section 3.1). This helps us study the impact of extra dimensions on the corpus model performance, and the top performance obtainable when the dimensionality is adapted to the corpus (Section 3.2). By looking at the trend of these performance curves, we determine the widely stable and well performing dimensionality. Once the optimal dimensionality is found, we compile test scores for the models with the optimal dimension count, varying the size of the corpus subsets (Section 3.3). This helps us quantify the corpus quantity-quality trade off in terms of the test performances.

3.1 Effect of Dimensionality

To explore the number of dimensions that is optimal for the different corpora and tasks, we run the behavioural and brain tests at different dimensionalities. We vary the number of SVD dimensions for the behavioural tests in steps of 25 over the range 25 to 1250 and we considered the SVD dimensions 75, 125, 250, 375, 500, 750 and 1000 for the neurosemantic decoding test (this less exhaustive search is due to the increased computational complexity of this task). Figure 2 shows the performance plots for the corpora across all six tests.

From Figure 2, we notice a general trend that all behavioural tests tend to improve as the number of dimensions increases. But, most of these tests flatten out after a particular point. In some case, at higher dimensions, we notice that the curve dips, presumably as noisy or irrelevant SVD dimensions are encountered. We find that the TOEFL test follows a very strong linear trend for most corpora. Extra dimensions seem to aid performance in this test. Books seems to have a substantial advantage in this test.

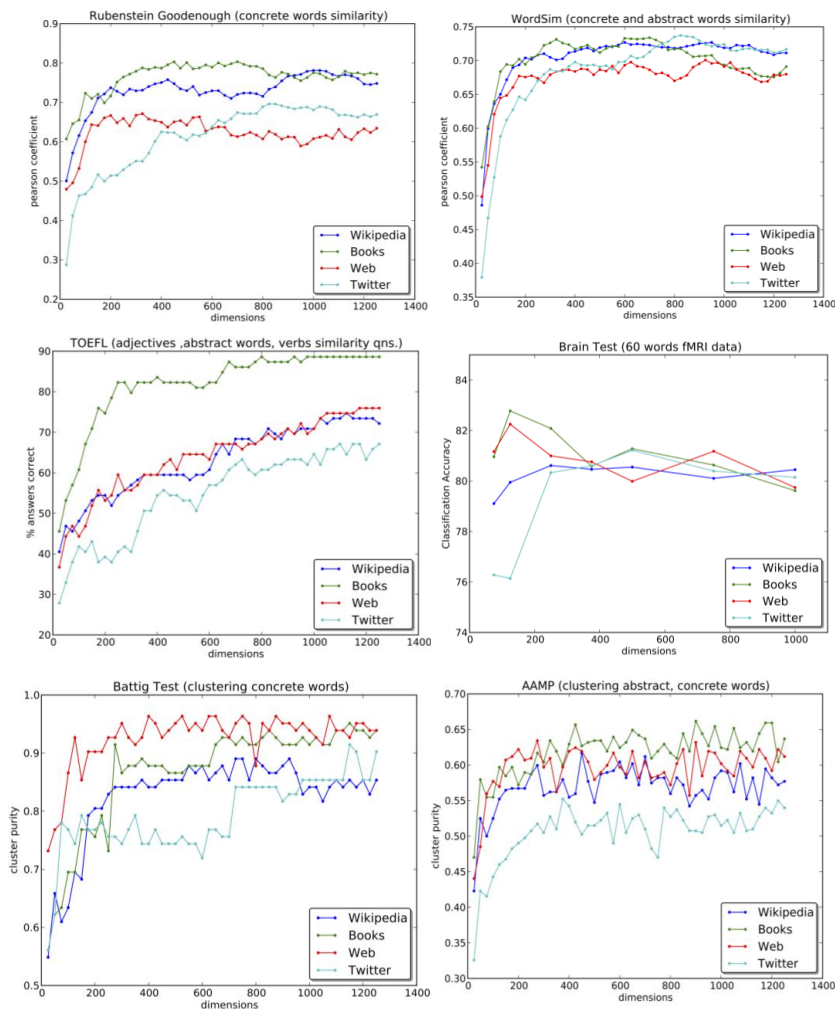


FIGURE 2 – Performance of corpus models over a range of SVD dimensions

The stability of the Twitter curves is lower than the others irrespective of the test. From the behavioural tests involving concrete nouns (Battig, R&G and WordSim), we notice that Twitter requires more SVD dimensions to attain peak performance, or reach a more stable score. The tests based on clustering (Battig and AAMP) appear to have unstable curves relative to the similarity and classification based tests. This has more to do with the nature of the tests than the corpus behaviour. The word clusters tend to vary greatly with dimensionality.

The Brain test is discriminating only at lower dimensions. Contrasting to the behavioural tests, extra dimensions do not affect the performance of the models noticeably. At lower dimensions, we find Books to perform best, and Twitter the worst. Model performances above 375 dimensions are all comparable. It is interesting that the Brain test peaks at such low dimensionality while tests like TOEFL and AAMP need a lot more dimensions. This could be because the top few SVD dimensions are more likely to be the important ones that encode the more common attributes (such as ‘living’ or ‘non-living’), which can help distinguish concrete nouns. On the other hand, a lot more dimensions are required to distinguish the more subtle differences between the TOEFL choices.

3.2 Peak Performances for Whole Corpora

As we saw in Figure 2, model performance is somewhat unstable, and varies in value and trend for different tests. To estimate the peak performance that is possible with each model, we aggregate over the top few points in the plots, reporting the average of the top 3 values rather than the very best value (Table 2). These results represent peak performances when dimensionality is tailored to each model/test pair. The best score among the corpora is highlighted in bold.

	Wikipedia	Twitter	Books	Web
Brain	0.81	0.81	0.82	0.82
Battig	0.89	0.94	0.94	0.96
R&G	0.78	0.69	0.80	0.66
WordSim	0.72	0.73	0.73	0.69
AAMP	0.61	0.54	0.66	0.63
TOEFL	0.74	0.67	0.89	0.76

TABLE 2 – Average of the top three scores over a range of dimensions

Overall, we find the performance of Books to be superior to the other corpora. It scores higher than the others in everything but the Battig test on which it is very close to the higher value achieved by the Web corpus. The Books model also has a substantial lead in the TOEFL and AAMP tests. This may be due to an advantage in capturing the meaning of adjectives, abstract nouns and verbs better. We explore the reasons for this performance gap later (Section 4). On the other hand, the Twitter model seems to perform very poorly over these two tests. Although the performances in the Brain test are very much similar, they exhibit the general performance trend observed in the other tests. As noted by Levy and Bullinara (2012) and Murphy et al (2012), the Brain test appears to have a performance ceiling, possibly due to noisiness in data. On close inspection of the Table, a trend emerges that Books is the best corpus model, followed by Web and Wikipedia, followed closely by Twitter.

3.3 Effect of Corpus Size

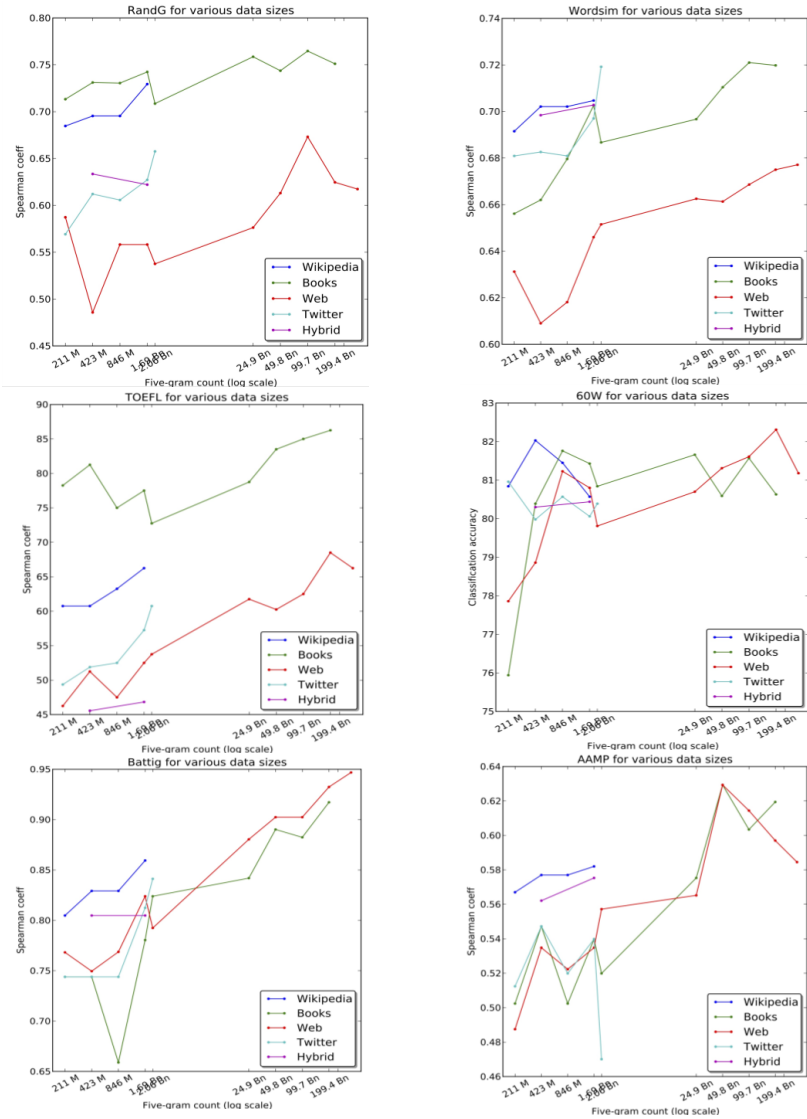


FIGURE 3 – Performance of corpus models over a range of data sizes

The corpus models analysed in the previous sections are on corpora of different sizes. Although it helps us establish trends in peak performance, this is not a fair comparison of the corpus types, given that there is enormous disparity in their original sizes (Table 1). To be able to compare the quality of the semantic models across corpora, we need to build them from corpora of similar sizes, at a dimensionality that is favourable to all these corpus types. From Figure 2 we notice that the performance for all corpora are stable and close to their peak values at 750 SVD dimensions. Although 750 dimensions may not be the most optimal dimensionality for these corpus models at different corpus sizes, we keep the dimensionality constant at 750, and vary the data sizes the models are learnt from. We under-sample (by random selection) the four different corpora to the 5-gram sizes of 12.5% Wiki, 25% Wiki, 50% Wiki, 100% Wiki, 100% Twitter, 12.5% Books, 50% Books and 100% Books. Also, at the sizes 50% Wiki and 100% Wiki we generate Hybrid corpora that have equal proportions of the randomly sampled five-grams from Wikipedia, Books, Web and Twitter. Figure 3 contains the performance plots of the corpus types across the above-mentioned data sizes, at 750 SVD dimensions. These new models are built similar to the old ones, as described in Section 2.2.

In many cases, we find an approximately log-linear trend in the performance with the corpus size for the behavioral test. We find models at smaller data sizes to be less stable, with more deviations from log-linearity. Although averaging scores over multiple randomly drawn samples may give a better approximation, it is extremely expensive and does not guarantee any bounds on the approximation. Regardless, all the original corpora considered are random samples of text of that nature to begin with. For this reason, we believe this approach is satisfactory to draw broad conclusions on the effects of corpus size, even if there exist local deviations from linearity.

In the behavioral tests, we notice that at smaller sizes, the Books and Web models generally have a considerable drop in performance. Wiki outperforms the others or achieves competitive scores at data sizes within 1.7 billion tokens. We notice that Books still retains its advantage in the TOEFL test at all data sizes. A similar advantage is observed to some extent for the Web corpus in the Battig test. The performance curves for Wiki are the most stable across all behavioral tests. The Brain test, unlike the behavioral tests does not seem to be affected by the sizes above 423 million five-grams. The curves stay flat across larger data sizes, with little increase in accuracies. Performance of the Hybrid set varies relative to the component corpora from which it is assembled. In some cases it is close to peak performance, but in others it underperforms considerably. This suggests that there is not beneficial complementarity among the corpora.

4 Discussion

4.1 What is the quality-size trade off?

Although we can see in detail how the corpus models perform in the six tests at different data sizes, we are primarily interested in how the corpora compare against each other. To understand the general quality of the corpus types, which is some function of the performances in the six tests, we compare them by their performances at a particular data size. We assign a rank to every result in the plot for a particular test, based on the score. Once this is done for all the tests, we compile the average rank for a corpus across all tests scores for a particular corpus size, which serves as a summarization of the 6 plots in Figure 3. Figure 4 shows the ranked model plots we obtain over different corpus sizes.

Corpus quality does have a considerable impact on the model performance. Although we see that the more the data, the better the performance, we clearly see that a corpus of high quality at a small size can perform better than a corpus of poor quality that is many orders of magnitude larger. At all corpus sizes up to 1.7 billion five-grams, Wikipedia is the best choice. The next best option in this size range is Books. Performance of Twitter and Web is comparable, although at very small sizes, Twitter performs better. The Hybrid corpus does not give us any advantage over the models that can be built from the constituent corpora. Above 2 billion five-grams, Books is the best choice, followed by Web.

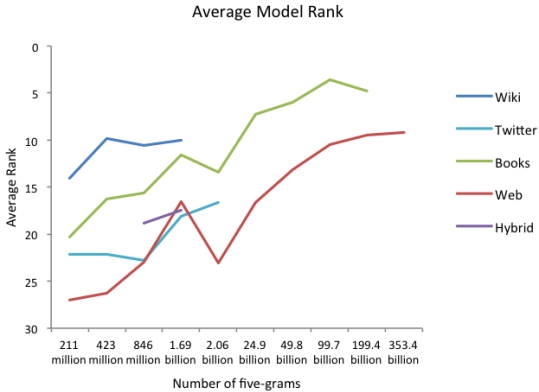


FIGURE 4 – Average rank over corpus size (in five-grams)

4.2 What is the quality-dimensionality trade off?

Working with a large number of dimensions can be expensive in certain applications. To determine the optimal number of dimensions for the corpora, we plot average corpus ranks across SVD dimensionalities (Figure 5) for the models built from the original corpus sizes. The response to dimensionality is not considerably different among these corpora. All of them perform better with extra dimensions up till a point after which they either fall in performance or flatten out. Although, Twitter in particular needs more dimensions than the other corpora to reach a similar rank. Also, Wikipedia tends to benefit more from extra dimensions than Web does.

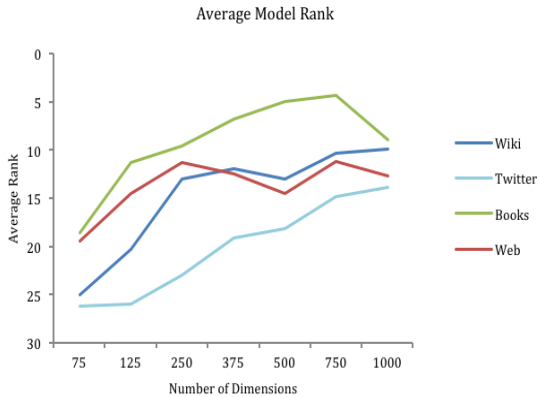


FIGURE 5 – Average rank across SVD dimensions

4.3 How different are the corpus models ?

To try to explain why we see these differences, we can explore whether the information encoded in the semantic models are in fact substantially different. Here we perform a follow-on analysis that measures the informational overlap between the corpus-derived models. We use the method introduced by Murphy et al (2012) to measure how much a model can explain the information contained by another model with the same vocabulary. We use the semantic models of 750 dimensions learnt from the full sized corpora. Besides analyzing the overlap for the common vocabulary among the corpus models, we also perform analyses for selected concrete nouns, abstract nouns, adjectives and verbs. From the MRC Psycholinguistic Norms (Coltheart.M, 1981), we first select nouns with the top 1000 concreteness score for the concrete-nouns list and those with the least 1000 concreteness scores for the abstract nouns list. From the American National Corpus (ANC) (Macleod et al. 2002) we include the top 1000 words that are adjectives 80% or more of the times into the adjectives list. We include the top 600 ANC words that are verbs 40% or more of the times into the verbs list. After calculating the information overlap values, we create a cosine similarity based isomap of the corpus-types for the five word groups by reducing the 4 dimensions of information overlap (with every corpus-type) down to 2 dimensions (Figure 6). In these corpus maps, the lesser the distance between two corpora, the more the common-information that is present, the more their semantic similarity.

As a general pattern, we notice that Twitter is the most semantically dissimilar among all corpora. This might be explained by its minimal lexical overlap and distinct language use. In the common-words and abstract-nouns plots, Web is equidistant from the other corpora. This is interesting since we assume the general-coverage Web to have a mix of the characteristics of the others. We find that the semantic information contained by the corpora for verb vectors has very high similarity. i.e. all corpora model verbs in a similar way. There is far lesser information overlap for adjectives, with Books and Twitter containing very different information compared to Wiki and Web. Books encodes very dissimilar information for abstract nouns as well. This could explain its differentiating performance in the TOEFL test.

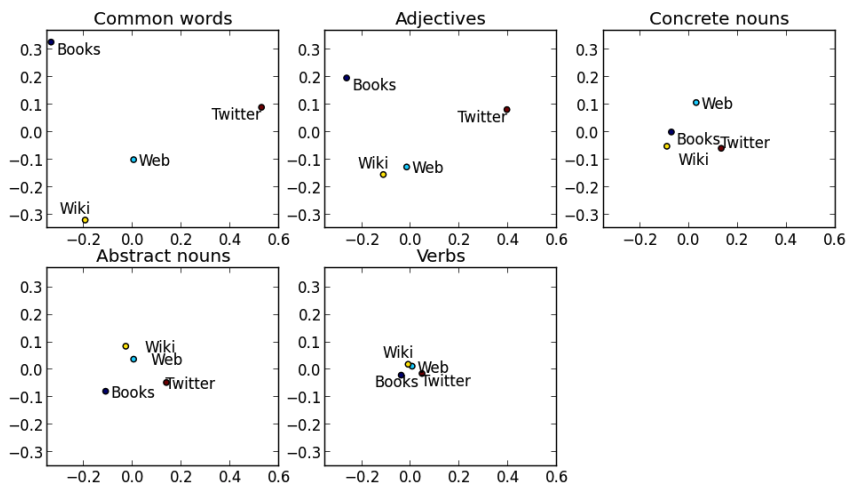


FIGURE 6 – Isomapped Corpora for different word-groups

4.4 Conclusions and future work

Given the wide array of corpus choices to build dictionaries based on distributional semantics, and their ubiquity, it is important to understand the contribution of corpus size and quality. From our experiments, it is evident that corpus choice does matter. Massive quantity is required to match the quality advantage. It is clear that Wikipedia is the corpus of choice for the data size in which it is available. The next most competitive corpus, Google-Books, must be an order of magnitude larger than Wikipedia before it can provide superior performance; and Google-Web must be two order of magnitude larger to match Wikipedia.

We speculate that the impressive performance of Wikipedia can be attributed to the balance in topics and cleanliness. While Books, a corpus roughly half the size of Web, is not as carefully balanced by topics, it presumably draws its advantage from its cleanliness and superior use of the language. On the other hand, the Twitter and Web do not exhibit any of these characteristics. Although the tweets have been heavily pre-filtered for our experiments, they probably still suffer from the colloquial nature of text, imbalance in topics and high rate of lexical errors. Web text also suffers from formatting errors, informal language use and imbalance in topics. These characteristics may have hampered their performance.

The advantage of quality over quantity for modeling word meaning, and the distinguishing performance of Wikipedia is a very interesting since the open-source encyclopedia is available in many languages at considerable sizes. As a next phase of this research, we plan to perform this analysis on similar corpora of other languages to study the generalizability of these results. We also plan to study the impact of different model types (directional, part of speech, dependency etc.) on corpus-derived model performance.

References

- Almuhareb, A., Poesio, M., (2004), Attribute-based and value-based clustering: an evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.
- Battig, W.F., Montague, W.E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. In *Journal of Experimental Psychology Monograph*, 80:1-45.
- Bellegarda, J.R. (2000). Large vocabulary speech recognition with multispan statistical language models. In *Proceedings of the IEEE Transactions on Speech and Audio Processing*, Jan 2000
- Brants, T., Franz, A.(2006). Web 1T 5-gram Version 1. In *Linguistic Data Consortium*, Philadelphia, PA, USA.
- Bullinaria, J.A. and Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. In *Journal of Behavior Research Methods*, 2007.
- Coltheart, M. (1981b). The MRC Psycholinguistic Database. In *Quarterly Journal of Experimental Psychology*, 33A, 497-505.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppim, E. (2002). Placing Search in Context: The Concept Revisited, In *Proceedings of the ACM Transactions on Information Systems*, 20(1):116-131.
- Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning*. In *Volume 18 of Springer Series in Statistics*. Springer, 5th edition.
- Karypis, G. (2002). CLUTO: a software package for clustering high Dimensional data sets. *University of Minnesota, Dept. of Computer Science*.
- Kilgariff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 97-233.
- Kilgariff, A. (2012). Getting To Know Your Corpus. *Lecture Notes In Computer Science*, Sojka, P., Horak, A., Kopecek, I., Pala, K. (eds). Springer.
- Landauer, Thomas, K. and Dumais, Susan T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211-240.
- Leacock C., Towell, G., Voorhees, E. (1993) Corpus-based statistical sense resolution, In *Proceedings of the Workshop on Human Language Technology*, Pages 260—265.
- Levy, J.P., & Bullinaria, J.A. (2012). Using enriched semantic representations in predictions of human brain activity. In *Proceedings of Connectionist Models of Neurocognition and Emergent Behavior: From Theory to Applications*, 292-308. Singapore: World Scientific, 2012.

Lin, D., Pantel, P. (2001). DIRT – discovery of inference rules from text. *Proceedings of the seventh ACM SIGKDD-International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA.

Robert L., Vladislav, D.V., Alex, G., Wayne D. G. (2007) Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness. In *Proceedings of the Eighth International Conference on Cognitive Modeling* 279–284.

Rubenstein, H., Goodenough, J. (1965). Contextual correlates of synonymy. In *Journal Commun. ACM* 8(10): 627-633.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. In *Journal Behavior Research Methods, Instruments, and Computers*, 28:203–208.

Macleod, C., Grishman, R. (2002). The American National Corpus: Standardized Resources for American English. In *Proceedings of 2nd Language Resources and Evaluation Conference (LREC)*, Athens, Greece.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320:1191–1195.

Murphy, B., Korhonen, A., Chang, K. K.-M., editors (2010). In *Proceedings of the Workshop on Computational Neurolinguistics, NAACL-HLT*, Los Angeles, CA, USA.

Murphy, B., Talukdar, P., and Mitchell, T. (2012). Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, QC, Canada.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. and Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington, DC, USA.

Rapp, R. (2003). Word Sense Discovery Based on Sense Descriptor Dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.

Sahlgren, M. (2006). The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. *PhD dissertation, Department of Linguistics, Stockholm University*.