# FAST: A carefully sampled and cognitively motivated dataset for distributional semantic evaluation

Stefan Evert Computational Linguistics Group FAU Erlangen-Nürnberg stefan.evert@fau.de

## Abstract

What is the first word that comes to your mind when you hear giraffe, or damsel, or freedom? Such free associations contain a huge amount of information on the mental representations of the corresponding concepts, and are thus an extremely valuable testbed for the evaluation of semantic representations extracted from corpora. In this paper, we present FAST (Free ASsociation Tasks), a free association dataset for English rigorously sampled from two standard free association norms collections (the Edinburgh Associative Thesaurus and the University of South Florida Free Association Norms), discuss two evaluation tasks, and provide baseline results. In parallel, we discuss methodological considerations concerning the desiderata for a proper evaluation of semantic representations.

# 1 Introduction

Assessing the performance of a distributional semantic model (DSM), be it a count model or one of the numerous and popular neural embeddings, has never been a straightforward endeavour. It is becoming an increasingly pressing issue due to the 'black box' nature of word embeddings and their aleatoric, often irreproducible training. We witness a proliferation of tasks and a growing dissatisfaction within the community: what are we modeling, really? Are models learning to handle a specific dataset, or a task, or do they really abstract *semantic* knowledge from text? This has lead to a shift towards evaluation methods capable of capturing human-like generalization (Linzen, 2020).

DSM evaluation with standard datasets is affected by two major problems. The first issue is the possibility that DSMs may just exploit contingent properties of a task. For example, they could be learning to identify random controls instead of capturing the intended semantic relation (Evert, 2016). Even with carefully chosen distractors, models may just be picking up contingent features of the experimental items: Levy et al. (2015b) observe that

Gabriella Lapesa Institute for Natural Language Processing University of Stuttgart lapesaga@ims.uni-stuttgart.de

many hypernym classifiers merely learn to recognize words that are 'typical' hypernyms, independently of the stimulus items (i.e. corresponding hyponyms). The second issue is dataset size, since small test sets inevitably lead to overfitting if many different model architectures or parameter settings are evaluated. Take the TOEFL synonym task, where a single item is worth 0.8 accuracy points, as an extreme example: Bullinaria and Levy (2012) achieve 100% accuracy after testing a wide range of DSM parameters. It is also true, however, that large datasets often lack in design quality (e.g. by using random controls, cf. above).

In this paper, we present Free ASsociation Tasks (FAST), a new evaluation dataset for English lexical semantics designed to satisfy the desiderata highlighted above: (a) compatibility with humanlike generalization, (b) quality of distractors, and (c) dataset size. We address (a) by focussing on a cognitive modeling task, namely free associations (which word comes to your mind when you hear giraffe?). Free association norms are cognitively motivated (hence semantically plausible) and usually very large, covering thousands of stimuli. This allows for a better selection of distractor items (which makes the task more difficult but also more enlightening), addressing both (b) and (c). As a bonus, we do not collect yet another dataset; instead, we carefully sample from existing collections and define two different evaluation tasks, providing extensive baseline results for both.

The first task we propose is a **multiple-choice task**:given a stimulus (e.g., *giraffe*) decide which of three candidates (*neck*, *apple*, *wine*) is its most frequent associate. It is in this task that we exploit the size of the underlying collection to collect better distractors: we don't sample them randomly, but rely on frequency criteria instead. Distractors are selected to match the frequency band of the correct candidate (*neck*), in order to prevent the models from relying on frequency effects. The second task

we propose is an open-vocabulary lexical access task: given a stimulus word (e.g., giraffe), select its most frequent associate from a large vocabulary of candidate words (consisting of the top associates for all stimuli in the data set). Advantages of the multiple-choice task are that it is easier to implement, less computationally expensive, lends itself to the application of machine-learning techniques as a straightforward classification problem, and has an obvious evaluation metric (accuracy). The open vocabulary task is cognitively more realistic and matches the experimental setting underlying free association norms, but it is computationally expensive (because of the larger number of candidates to be considered), poses a harder machine-learning problem (learning to rank), and suitable evaluation metrics are less obvious (see Sec. 4.1).

This paper makes the following contributions: we release FAST and propose two new evaluation tasks based on it; we report preliminary modelling results as a baseline for further work.

# 2 Related work

Evaluation in terms of cognitive modeling is challenging from many points of view. First, the modeled data are typically continuous (reaction times or EEG signal, see Mandera et al. (2017) and Hollenstein et al. (2019) for an overview) or, when categorical (free association norms), much less constrained compared to the data used in standard DSM evaluation tasks because the models have to handle a large and diverse vocabulary. Second, a speaker's behaviour results from the interplay of many factors, some of which we expect to be conflated in distributional meaning representations and intertwined with 'true' lexical semantics (e.g, ambiguity, culture-specific information), others to be completely absent (e.g., individual differences between speakers). Third, cognitive modeling datasets typically have not been designed to evaluate corpusbased models and thus need to be further manipulated (annotated, adapted, sampled), bringing in additional design assumptions. To sum up, while quantifying DSM performance is straightforward in standard tasks, it is less so in the case of cognitive modeling tasks.

#### 2.1 Free Associations

Free associations ("the first word that comes to mind when you hear  $\dots$ "), are considered to be a cue into the organization of the mental lexicon.

The nature of the involved cognitive processes has always been debated: while earlier theories considered free associations as the result of learning by contiguity (James, 1890), later theories accounted for them in terms of symbolic processes and complex semantic structures (Clark, 1970). Annotation studies resulted in the characterization of free associates as a mixture of syntagmatic relations, which hold between contiguous word, and paradigmatic relations, which hold between semantically related words (Brown and Berko, 1960; Fitzpatrick, 2007).

In particular, Brown and Berko (1960) observed that in a typical free association setup, 74% of the responses of adults are paradigmatic while 72% of the responses of first grade children are syntagmatic. Fitzpatrick (2007) conducted a free association experiment on 100 stimuli and 60 subjects and manually classified the associates, identifying the following categories: Consecutive xy collocations: *significant–other*; Defining synonyms: *significant– important*; Conceptual associations: *coordination– driving*; Consecutive yz collocations: *instance– first*; Context-dependent synonyms: *label–name*; Lexical set: *pet–dog*.

# 2.2 Corpus-based modeling

Based on a free association dataset, two types of evaluation tasks can be designed: in the 'forward' task, the corpus-based model has to generate the first free associate (i.e. the one produced by the highest number of subjects) for a specific stimulus  $(cat \rightarrow ?, significant \rightarrow ?)$ ; in the 'reverse' task, the model is shown a number of response words and has to guess the corresponding stimulus (? $\rightarrow away$ , minded, gone, present, ill). A problem of this evaluation setup is the unrestricted set of possible responses in combination with a discrete association task, which requires the algorithm to pick exactly the right answer out of tens of thousands of possible responses. This feature makes the task much more difficult than the multiple-choice tasks often used to evaluate DSMs. Additionally, free association datasets have been employed in a multiple-choice setting, in which the model has to pick the first associate to a stimulus word given a small number of candidate responses.

A multiple-choice task derived from Edinburgh Associative Thesaurus (EAT) appeared in the ESS-LLI 2008 shared task.<sup>1</sup> Predictions based on firstorder co-occurrence turned out to be much better

wordspace.collocations.de/fa2008.html

than those from second-order DSMs, in line with previous findings by Rapp (2002) and Wettler et al. (2005). A similar picture emerges from studies on the reverse (or "multiword") free association task: models based on first-order co-occurrence outperform models based on vector similarity. This superiority, however, has not been demonstrated in a direct comparison. Results were obtained by studies with different features and goals (see Rapp (2014) for a review; see Griffiths et al. (2007) for an evaluation of models based on Latent Semantic Analysis). Successful studies on the reverse association task include Rapp (2013, 2014), as well as the CogALex 2014 shared task (Rapp and Zock, 2014) on 2000 stimuli from EAT. This task proved to be very challenging: the winner, which used first-order statistics to re-rank the output of a DSM, only achieved 35% accuracy (Ghosh et al., 2015).

# 3 Dataset

The starting point for FAST are the two largest collections of free association norms available: the Edinburgh Associative Thesaurus (EAT, 8210 stimuli, 100 subjects; Kiss et al., 1973) and the University of South Florida Free Association Norms (USF, 5019 stimuli, 6000 subjects; Nelson et al., 2004).

**Pre-processing** All items in EAT and USF were POS-tagged using frequency information from the large, publicly available Web corpus ENCOW 2014<sup>2</sup> to guess the most probable POS tag for each word form out of context. POS tags are not employed in the evaluation experiments conducted in this paper, but we believe that guessing POS based on frequency is cognitively plausible for out-ofcontext tasks and distribute POS annotation with the official release, to support further evaluation with POS-disambiguated DSM representations (e.g. **Baroni and Lenci**, 2010). We chose ENCOW because it is publicly available, large (10 bln words) and thus suitable for the extraction of reliable cooccurrence estimates and vector representations.

FAST contains lemmatized items. This choice has practical reasons – many DSMs and other evaluation datasets are also lemmatized – and is supported by the literature. Bullinaria and Levy (2012) compare DSMs built from raw, stemmed and lemmatized data. In a selection of semantic similarity tasks, a slight advantage is found for the lemmatized and stemmed models over the raw ones (with variations due to the different tasks, and to the interaction with other parameters). For a different selection of tasks, Kiela and Clark (2014) report best performance for stemmed data vs. inflected, lemmatized, POS- and CCG-tagged.<sup>3</sup> Lapesa and Evert (2013) obtain similar results for modeling semantic priming. We carried out lemmatization with morpha, a robust morphological analyzer.<sup>4</sup> Unknown words were lemmatized based on their POS tag. All items in USF and EAT were annotated with their ENCOW lemma frequency (using the same lemmatization), as a basis for our sampling procedure.

**Dataset compilation** The compilation procedure of the FAST dataset is based on the following assumptions: (a) words that were produced only by a single subject in response to a stimulus can be used as plausible distractors in a multiple-choice task; (b) first responses to other stimuli (i.e. typical first associates) are good 'implausible' distractors; (c) the large pool of filler candidates allows controlling for frequency (as DSM similarities often show frequency bias). Multiwords,<sup>5</sup> numbers, closed-class words, and words not occurring in ENCOW were discarded. For each remaining stimulus in EAT and USF we selected:

- **FIRST**: the most frequent associate response;
- HAPAX: a response generated only once for the stimulus (twice for USF, which omits responses with f = 1);<sup>6</sup> among several HAPAX candidates, we picked the one whose lemma frequency was closest to that of FIRST;
- **RANDOM**: a randomly selected response from the top 25% associates *of another stimulus* (and produced ≥ 5 times); if possible we
  - matched lemma frequency of RANDOM and FIRST, and
  - used each RANDOM only once.

<sup>&</sup>lt;sup>2</sup>corporafromtheweb.org/encow14/

<sup>&</sup>lt;sup>3</sup>CCG tags contain lexical categories based on the Combinatory Categorial Grammar (Steedman, 2000) which can be thought of as finer-grained POS-tags.

<sup>&</sup>lt;sup>4</sup>users.sussex.ac.uk/~johnca/morph.html

<sup>&</sup>lt;sup>5</sup>This choice is due to practical reasons: most pre-compiled DSMs and off-the-shelf embeddings do not cover multiwords. Since there are very few such entries in USF (and a moderate number in EAT), the additional pre-processing steps needed for including them in our baseline experiments would have been unjustified.

<sup>&</sup>lt;sup>6</sup>Nelson et al. (2004) argue in their introductory essay that "idiosyncratic responses given by a single participant would tend to be 'off the wall'."

These items were randomly divided into training and test sets, resulting in 3774 + 3836 items for EAT and 2360 + 2359 items for USF.

FAST has been publicly released in the Open Science Framework, along with complete RMarkdown replication scripts for the baseline experiments. It is available from https://osf.io/cd8ar/.

# 4 Experimental setup

# 4.1 Tasks

We propose two evaluation tasks and present baseline results for the EAT and USF test subsets.

The first is a **multiple-choice** task. Given a stimulus and a  $<\underline{FIRST}$ , HAPAX, RANDOM> triple (e.g., from FAST, the triple  $<\underline{receive}$ , love, soul> for the stimulus accept), the goal is to determine which of the candidates is FIRST. Performance is quantified as percent accuracy, and with three candidates the random baseline is 33.3%.

The second task is an **open-vocabulary lexical access** task: given a stimulus (e.g., *accept*), predict the FIRST associate (*receive*) out of a set of candidates which consists of all FIRST responses in the respective subset (USF or EAT). This task is operationalised as a ranking of the entire candidate set for each stimulus (candidate set size: USF=1197; EAT=1633).<sup>7</sup>

For the open-vocabulary lexical access task, we propose two evaluation metrics: Soft accuracy and Log rank. Soft accuracy averages over reciprocal rank (1/r) of the true FIRST associate in the candidate set, expressed as a percentage. This metric corresponds to a 'soft' version of accuracy that awards partial points if the correct choice is not in first rank, but close. Log rank computes the geometric mean of rank (r) across all stimuli. It is more informative than soft accuracy for task items that the models don't solve well, because it distinguishes more clearly between a correct answer at rank 50 or rank 500. Since the arithmetic mean would penalize single items with poor ranks (e.g. one at rank 1000) too much, we compute the geometric mean of the ranks. We refer to this as log rank  $(\log r)$  because it corresponds to averaging ranks on a logarithmic scale; but note that the values reported are not logarithms. This metric is more suitable for models that perform reasonably well (with correct response often among first 100 ranks), but rarely get the correct answer (and hence score low on (soft) accuracy).

<sup>7</sup>NB: Some words are first responses for multiple stimuli.

In the open-vocabulary task, the baseline corresponds to a random ranking of the candidate vocabulary. Mathematically, we compute the expected values of soft accuracy and log rank for a uniform distribution over the possible ranks of the correct choice, and we are aware that this is a weak baseline.

Baselines depend on the size of the candidate vocabulary and are thus different for USF and EAT: soft accuracy: USF=0.64% and EAT=0.49%; log rank: USF=442.0 and EAT=602.4.

#### 4.2 Models

We experimented with the following models, created from 8.5G tokens of the ENCOW 2014 Web corpus unless otherwise specified:

First order co-occurrence (collocations), based on a symmetric span of 2 vs. 10 words and quantified by conditional probability ( $P(w_2|w_1)$ ), logtransformed G<sup>2</sup>, PPMI, and MI<sup>2</sup> (Evert, 2008).

*Count DSMs (second order)*, based on a symmetric span of 2 vs. 10 words and SVD dimensionality reduction, with other parameter settings as in Lapesa and Evert (2014).<sup>8</sup> We experimented with one further parameter, *Caron P* (Bullinaria and Levy, 2012; Levy et al., 2015a). Standard SVD-reduced models were compared with their P = 0 counterpart; this operation has the effect of scaling down the prominence of the first SVD dimensions in the semantic representation and tailoring them towards paradigmatic relations.<sup>9</sup>

<sup>9</sup>Singular value decomposition (SVD) factorizes the cooccurrence matrix  $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T$ , where  $\Sigma$  is a diagonal matrix of ordered singular values  $\sigma_1 \ge \sigma_2 \ge \ldots$  showing how much of the variance is captured by each latent dimension. The dimensionality-reduced representation is thus obtained by truncating the factorization to the first r (i.e. most important) components:  $\mathbf{U}_r \Sigma_r$ . Caron (2001) found out that raising the singular values to a power P < 1 and thus reducing the prominence of the first components improved the effect of SVD on LSA representations. This result was later replicated for termterm models (Bullinaria and Levy, 2012; Levy et al., 2015a). Optimal results are often obtained for P = 0 (i.e. using  $\mathbf{U}_r$  as the dimensionality-reduced representation), which equalizes the contributions of all dimensions. The effect is comparable

<sup>&</sup>lt;sup>8</sup>Target terms: All POS-disambiguated lemmas from the FAST data set combined with a standard vocabulary of frequent words, 58k terms in total. Context dimensions: same 58k POS-disambiguated lemmas as target terms. Feature scoring: sparse simple-log likelihood with an additional log transformation. Similarity computation: cosine similarity (using angle as a distance metric). SVD-reduced to 1000 latent dimensions (note that the general settings in Lapesa and Evert (2014) would recommend to discard the first 50 dimensions, but we experiment with Caron *P* instead; also note that they recommend 500 SVD dimensions, but their experiments also show that increasing SVD dimensions is always advantageous and never detrimental).

Combined first- and second-order, based on the harmonic mean of neighbour ranks. Since cosine similarity (second-order) and association measures (first-order) are on entirely different scales, it would be difficult to combine them into a single composite score for a given word pair (A, B). We therefore use neighour rank, i.e. the rank of B among the nearest neighbours of A, as an index of distributional relatednes, following Lapesa and Evert (2014). For the first-order data, we order all words in the vocabulary by decreasing association score and take the rank of B in this ordering as its neighbour rank (Michelbacher et al., 2011). We consider the harmonic mean a better strategy for combining ranks than average, minimum or maximum because it is not dominated by extreme values. To exploit complementarity between different DSMs we a) combine a DSM known to perform well in paradigmatic tasks (with P = 0) with a first-order association known to perform well in collocation extraction (MI<sup>2</sup>), even if they are not the best individual models, and b) in addition to the span size of 2 vs. 10 words, we also experiment with a mixed span of size 2 for the DSM and of size 10 for first-order collocations (span mix in the result tables).

*Embeddings*: word2vec (Mikolov et al., 2013) trained on 100G tokens of Google News data,<sup>10</sup> GloVe (Pennington et al., 2014) trained on 6G tokens of wikipedia + newspapers or 42G tokens of Common Crawl (CC) Web data,<sup>11</sup> and FastText (Joulin et al., 2017) trained on 600G tokens of CC.

We did not include contextualized embeddings because free associations are inherently an out-ofcontext task. We are aware that type-level representations can be obtained (Bommasani et al., 2020), but this is not the primary purpose of contextualised models and it would imply a number of design choices without clear guidelines. Given that the purpose of this paper is to release the dataset and provide baselines, we decided to stick to wellestablished type-level models.

## 5 Results

Tables 1 and 2 report evaluation results on the multiple-choice and lexical access task, respec-

tively. Items not covered by a model (*miss*) are ignored for the evaluation metrics, so as not to give an unfair advantage to models built specifically for the task vocabulary.<sup>12</sup>

Overall, first-order models outperform DSMs in the multiple-choice task, while DSMs are better in the lexical access task, achieving optimal performance with larger spans (known to introduce some first-order information into the DSMs). Setting P = 0 improves lexical access only slightly, and it has a detrimental effect on the multiple-choice task (presumably as it discards first-order information, see also Lapesa and Evert, 2014).

Combining first-order (collocations) and secondorder (DSMs) information results in a further improvement, and a mixed span yelds the best performance in the multiple choice task on EAT. For the lexical access task, the combined model with a mixed span consistenly outperforms both the 2word and 10-word combinations, even though the improvement over the 10-word span is small, showing that our assumptions regarding the complementarity of the two models may not entirely be justified.

To further explore the complementarity between first-order (collocations) and second-order (DSMs), we visualize the correlation between neighbour ranks of the two individual models in Figures 1 (USF) and 2 (EAT). Each point corresponds to a test item, i.e. a pair of stimulus and FIRST response, with DSM (2-word window, P = 0) neighbour rank on the x-axis and first-order (10-word window MI<sup>2</sup>) neighbour rank on the y-axis. Note that both axes are logarithmic and jitter has been added to visualize the large number of items where both ranks are 1 (etc.). The plots show that the two models are often in agreement, i.e. both the DSM and the first-order model put the FIRST response among the top 5 ranks. Indeed, there seems less complementarity to be exploited than we would have assumed, i.e. points in the bottom right or top left quadrants of the plot, for which the two models disagree substantially.

For both tasks and both subsets, the best com-

to discarding the first SVD dimensions, which Lapesa et al. (2014) found beneficial for modeling paradigmatic relations.  $^{10}$ https://code.google.com/archive/p/

word2vec/ <sup>11</sup>https://nlp.stanford.edu/projects/

glove/

<sup>&</sup>lt;sup>12</sup>In the multiple-choice task, an item is considered "missing" if either the stimulus is not included in the model or none of the three candidate responses is. In the open-vocabulary task, an item is considered "missing" if either the stimulus or the correct response is not included in the model. The DSMs and first-order data were specifically compiled for the FAST data set, so they achieve full coverage. The FastText model uses subword embeddings to approximate unknown words, also resulting in full coverage.

| USF<br>acc | EAT<br>acc  |
|------------|---|
|            |   |
| 16 (11/7   | 81.78%  |
|            | 81.78%<br>78.62%  |
|            | 78.02 <i>%</i><br>82.46 <i>%</i>  |
|            | <b>82.40%</b><br>79.30%   |
| 70.39%     | 79.30%  |
| 73.76%     | 80.29%  |
| 71.98%     | 78.28%  |
| 68.80%     | 76.12%  |
| 73.51%     | 79.87%  |
| 77.58%     | 84.02%  |
| 77.83%     | 83.00%  |
| 73.80%     | 81.18%  |
| 78.64%     | 83.92%  |
| 77.66%     | 85.01%  |
| 80.58%     | 85.74%  |
| 80.46%     | 86.03%  |
| miss: 7    | miss: 119   |
| 76.11%     | 77.78%  |
| miss: 2    | miss: 44  |
| 76.71%     | 79.80%  |
|            | miss: 7   |
| 80.12%     | 81.72%  |
| 82.24%     | 83.97%  |
| miss: 7    | miss: 116   |
| 76.53%     | 79.25%  |
| miss: 2    | miss: 38  |
| 77.17%     | 81.07%  |
|            | miss: 2   |
| 80.75%     | 83.07%  |
| 83.21%     | 85.58%  |
|            | 71.98%<br>68.80%<br>73.51%<br>77.58%<br>77.83%<br>73.80%<br><b>78.64%</b><br>77.66%<br><b>80.58%</b><br>80.46%<br>miss: 7<br>76.11%<br>miss: 2<br>76.71%<br>80.12%<br><b>82.24%</b><br>miss: 7<br>76.53%<br>miss: 2<br>77.17%<br>80.75% |

Table 1: Multiple choice task (baseline: 33.33%). 1st block: DSMs; 2nd: first-order; 3rd: combined first-order and DSM; 4th/5th: embeddings (lemmatized vs. wordform). Bold: best model per block.

bined model is only outperformed by pre-trained FastText embeddings, which achieve best results in all tasks except multiple choice on EAT. However, these embeddings were trained on 600G tokens of data, whereas the DSMs and collocations are based on only 8.5G tokens (and outperform other embeddings trained on similar-sized or larger corpora). Another highly speculative explanation for the better performance of FastText is its use of subword embeddings, which may be sensitive to some rhyming effects that are are known to be present in free associations (Nelson et al., 2004).

Since all pre-trained embeddings are based on surface forms, we also evaluate them in the nonlemmatized task (*wf* in the tables). These results are reported for completeness, but are not directly comparable to the DSMs and collocations. It is not clear yet whether the better results in this setting are a feature of the embeddings (because the uninflected form might not reflect the typical usage

|                    |      | n = 2359<br>USF |      | n = 3836<br>EAT   |      |
|--------------------|------|-----------------|------|-------------------|------|
| model              | span | soft acc. lrank |      | soft acc. lrank   |      |
| DSM                | 2    | 41.54%          | 6.6  | 34.53%            | 9.9  |
| $DSM_{P=0}$        | 2    | 42.12%          | 7.6  | 34.67%            | 12.1 |
| DSM                | 10   | 42.01%          | 6.0  | 35.93%            | 9.1  |
| $\text{DSM}_{P=0}$ | 10   | 42.86%          | 7.1  | 35.68%            | 11.6 |
| $P(w_2 w_1)$       | 2    | 23.30%          | 15.4 | 14.77%            | 22.8 |
| $\log G^2$         | 2    | 30.22%          | 10.7 | 28.25%            | 13.3 |
| PPMI               | 2    | 31.48%          | 11.5 | 27.08%            | 15.5 |
| $MI^2$             | 2    | 32.79%          | 9.1  | 29.97%            | 11.6 |
| $P(w_2 w_1)$       | 10   | 22.34%          | 17.0 | 11.27%            | 27.1 |
| $\log G^2$         | 10   | 37.63%          | 6.6  | 34.13%            | 8.8  |
| PPMI               | 10   | 35.34%          | 8.2  | 29.29%            | 12.2 |
| $MI^2$             | 10   | 39.73%          | 6.2  | 34.01%            | 8.7  |
| Comb.              | 2    | 42.29%          | 5.5  | 37.54%            | 7.0  |
| Comb.              | 10   | 44.99%          | 4.8  | 39.48%            | 6.5  |
| Comb.              | mix  | 45.36%          | 4.8  | 39.48%            | 6.4  |
|                    |      | miss: 7         |      | miss: 119         |      |
| w2v                | -    | 38.98%          | 7.7  | 30.51%            | 14.8 |
|                    |      | miss: 2         |      | miss: 44          |      |
| GloVe              | -    | 39.22%          | 7.6  | 30.19%            | 13.8 |
|                    |      |                 |      | miss: 7           |      |
| GloVe CC           | -    | 44.01%          | 5.7  | 34.26%            | 10.5 |
| FastText CC        | -    | 51.00%          | 4.1  | 40.34%            | 7.2  |
|                    |      | miss:           |      | miss: 116         |      |
| w2v wf             | _    | 39.60%          | 7.4  | 31.29%            | 13.5 |
|                    |      | miss: 2         |      | miss: 3           |      |
| GloVe wf           | -    | 39.74%          | 7.3  | 31.07%            | 12.5 |
| GloVe CC wf        | _    | 44.50%          | 5.5  | miss: 2<br>34.91% | 9.7  |
| FastText CC wf     | _    | 51.43%          | 3.9  | 41.30%            | 6.4  |

Table 2: Open-vocabulary lexical access task. 1st block: DSMs; 2nd: first-order; 3rd: combined first-order and DSM; 4th/5th: embeddings (lemmatized vs. wordform). Bold: best model per block.

of the lemma in corpora) or of the free association task (because human subjects may have different associations for different inflected forms).

# 6 Conclusion

This paper introduced the FAST dataset and discussed extensive baseline results in two carefully designed tasks.

The results on the multiple-choice task showed that first-order co-occurrence models outperform count-based DSMs, due to the known prevalence of syntagmatic relations in free associations. Combining the predictions of co-occurrence models and DSMs turned out to be particularly advantageous, with results competitive to those achieved by the neural embeddings in particular for EAT.

In the lexical access task, DSMs are slightly superior to co-occurrence models, and the combination of first-order and second-order data again leads to a substantial improvement. Here, Fast-



Figure 1: Correlation between first-order and DSM neighbour ranks for pairs of stimulus and FIRST associate in the USF test set.

Text embeddings achieve the best results, but our combined model still outperforms all other neural embeddings.

Besides the integration of more recent datasets (De Deyne et al., 2018), future work should target more sophisticated methods for the combination of different models, e.g., using machine learning techniques rather than just taking the harmonic mean as in our baseline experiments, and should also explore type-level representations derived from contextualized embeddings.

As far as the interpretation of model predictions is concerned, FAST lends itself to a multifactorial meta-analysis based on, e.g., abstractness ratings, semantic classes and ortographic properties of the stimulus. Such rich information sources could also be integrated in the prediction step (e.g., constrain the system to predict associates at a similar level of abstractness to the stimulus).

## Acknowledgments

We are very grateful to the anonymous reviewers for their feedback. Gabriella Lapesa acknowledges funding by the Bundesministeriumfur Bildung und Forschung (BMBF) through the project E-DELIB (*Powering up E-DELIBeration: towards AI-supported moderation*).



Figure 2: Correlation between first-order and DSM neighbour ranks for pairs of stimulus and FIRST associate in the EAT test set.

## References

- Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–712.
- Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of ACL*, pages 4758–4781, Online.
- R. Brown and J. Berko. 1960. Word association and the acquisition of grammar. *Child Development*, 31:1–14.
- John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44(3):890–907.
- John Caron. 2001. Experiments with LSA scoring: Optimal rank and basis. In Michael W. Berry, editor, *Computational Information Retrieval*, pages 157– 169. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
- H.H. Clark. 1970. Word associations and linguistic theory. In J. Lyons, editor, *New horizons in linguistics*. Harmondsworth: Penguin.
- Simon De Deyne, Danielle Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "small world of words": English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51.
- Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics*. *An International Handbook*, chapter 58, pages 1212– 1248. Mouton de Gruyter, Berlin, New York.

- Stefan Evert. 2016. CogALex-V shared task: Mach5 a traditional DSM approach to semantic relatedness. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 92–97, Osaka, Japan.
- Tess Fitzpatrick. 2007. Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics*, 17(3).
- Urmi Ghosh, Sambhav Jain, and Soma Paul. 2015. A two-stage approach for computing associative responses to a set of stimulus words. In *Proceedings* of the 4th Workshop on Cognitive Aspects of the Lexicon,, pages 15–21, Dublin, Ireland.
- Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114:211–244.
- Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceed*ings of CoNLL, pages 538–549, Hong Kong, China.
- W James. 1890. *The principles of psychology*. New York: Dover.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of EACL*, pages 427–431, Valencia, Spain.
- Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30, Gothenburg, Sweden.
- G.R Kiss, C. Armstrong, Milroy, and J. Piper. 1973. An associative thesaurus of english and its computer analysis. In R. Beiley Aitken and N. Hamilton-Smith, editors, *The computer and literary studies*. Edinburgh University Pres.
- Gabriella Lapesa and Stefan Evert. 2013. Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013)*, pages 66–74.
- Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions* of the Association for Computational Linguistics, 2:531–545.
- Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics* (\*SEM 2014), pages 160–170, Dublin, Ireland.

- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015a. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015b. Do supervised distributional methods really learn lexical inference relations? In *Proceedings* of NAACL-HLT, pages 970–976, Denver, Colorado.
- Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of ACL*, pages 5210–5217, Online.
- Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57 – 78.
- Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, 7(2):245–276.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings.*
- 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*.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, Doha, Qatar.
- Reinhard Rapp. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of COLING*, Taipei, Taiwan.
- Reinhard Rapp. 2013. From stimulus to associations and back. In *Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science*.
- Reinhard Rapp. 2014. Corpus-based computation of reverse associations. In *Proceedings of LREC*, pages 1380–1386, Reykjavik, Iceland.
- Reinhard Rapp and Michael Zock. 2014. The CogALex-IV shared task on the lexical access problem. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 1–14, Dublin, Ireland.
- Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA, USA.
- Manfred Wettler, Reinhard Rapp, and Peter Sedlmeier. 2005. Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, 12(2–3):111–122.