Evaluating word embeddings with fMRI and eye-tracking

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Abstract

The workshop CfP assumes that downstream evaluation of word embeddings is impractical, and that a valid evaluation metric for pairs of word embeddings can be found. I argue below that if so, the only meaningful evaluation procedure is comparison with measures of *human word processing in the wild*. Such evaluation is non-trivial, but I present a practical procedure here, evaluating word embeddings as features in a multi-dimensional regression model predicting brain imaging or eye-tracking word-level aggregate statistics.

What's the meaning of embeddings? In order to decide how to evaluate word embeddings, we first need to decide what word embeddings are supposed to encode. If we assume that word embeddings are primarily representations of the *meaning* of words, it makes sense to consult lexical semantic theories.

Here's a very, very, very (very, ...) crude characterization of lexical semantics: Researchers disagree whether words are defined by their co-occurrences (Firth, 1957), the contexts in which they are used (Wittgenstein, 1953), how they are organized in the brain (Miller and Fellbaum, 1992), or the referents they denote in the real world (Montague, 1973). I realize this is a ridiculously simplistic reduction of modern lexical semantics, but I think it suffices for our discussion of how best to evaluate word embeddings.[1]

Any metrics here? From (one or more of) these theories we want to derive a valid evaluation metric. In my view, *a valid metric satisfies two principles: (i) that it measures what we want to measure (adequacy), and (ii) that it cannot easily be hacked.* What I mean by (i) is that we want word embeddings to capture the meaning of words; and by (ii), that the reason we want to play the evaluation game is because it isn't obvious what the meaning of a word is. If the meaning of a word was given directly by its character sequence, I would not be writing this paper, and this workshop would not have been proposed. The question then is, do any of the four theories above provide us with a valid metric for the general quality of word embeddings?

Below, I argue that none of the four theories leave us with fully valid evaluation metrics, except maybe COGNITIVE LEXICAL SEMANTICS. I suggest evaluating embeddings by direct comparison with brain-imaging and eye-tracking data rather than word association norms, as an alternative approach to COGNITIVE LEXICAL SEMANTICS. I show that state-of-the-art embeddings correlate poorly with such data, but argue that this is nevertheless the only valid metric left on the table, if downstream evaluation is not an option – and that, practically, we can evaluate embeddings by the error of a multi-dimensional regression model predicting brain imaging or eye-tracking data using the embeddings as features.

Co-occurrence theory In CO-OCCURRENCE THEORY, the meaning of a word is defined by its co-occurrences with other words – e.g., the meaning of *big* is given by its co-occurrence with words such as *house* and *small*, i.e., its value in a co-occurrence matrix. Word embeddings should therefore predict lexical co-occurrences, which can be evaluated in terms of perplexity or word error rate. This was how embeddings were evaluated in the early papers, e.g., (Mikolov et al., 2010). But note that constructing co-occurrence matrices is also an integral part of standard approaches to *inducing* embeddings (Levy et al., 2015). In

[1] See the discussion in the last paragraph.

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 116–121,
Berlin, Germany, August 12, 2016.

fact for any definition of a word's *company*, we can built co-occurrence matrices tailored to maximize our objective. The associated metrics can thus be "hacked" in the sense that the encodings used for evaluation, can also be used for induction. Just like with other intrinsic evaluation metrics in unsupervised learning, co-occurrence-based evaluation easily bites its own tail. As soon as we have defined a word's *company*, the quality of the embeddings depends solely on the quality of the data. The evaluation strategy becomes the induction strategy, and the validity of the embeddings is by postulate, not by evidence. In other words, the metric can be hacked. Note that whether such a metric is *adequate* (measuring meaning) remains an open question.

Sprachspiel theory In SPRACHSPIEL THEORY, the meaning of a word is defined by its usage, i.e., the situations in which it is used. In Wittgenstein's words, *only someone who already knows how to do something with it, can significantly ask a name.* Obviously, it is hard to parameterize contexts, but explicit semantic analysis (Gabrilovich and Markovitch, 2009) presents a simple approximation, e.g., thinking of Wikipedia sites as contexts. Learning word representations from inverted indexings of Wikipedia is encoding a situational lexical semantics, albeit in a somewhat simplistic way. The meaning of *big*, for example, is defined by the Wikipedia entries it occurs in, i.e., its value in a term-document (or term-topic or term-frame or ...) matrix. The question then is: How well do our embeddings distinguish between different contexts? See earlier work on using embeddings for document classification, for example. However, such an encoding has also been proposed as an approach to *inducing* embeddings (Søgaard et al., 2015). While this proposal adopts a specific encoding of term-document matrices, similar encodings can be built for any definition of a *Sprachspiel*. Any such metric can thus be "hacked" or build into the model, directly. Note, again, that whether such a metric is *adequate* (measuring meaning) remains an open question.

Cognitive lexical semantics How well does our embeddings align with our mental representations of words? Obviously, we do not have direct access to our mental representations, and most researchers have relied on word associations norms instead.[2] In matrix terms, COGNITIVE LEXICAL SEMANTICS defines the meaning of a word as a vector over vertices in an ontology or a mental lexicon. The hypothesis is that our mental lexicon is organized as a undirected, colored, weighted network, and the meaning of words are defined by the edges connecting them to other words. The meaning of *big*, for example, is in a synonym relation with *large*, an antonym of *small*, etc. Such networks are typically informed by word association norms and corpus linguistic evidence. Using Wordnets for evaluating word embeddings was recently proposed by Tsvetkov et al. (2015).

However, again, Faruqui and Dyer (2015) recently proposed this as a learning strategy, encoding words by their occurrence in Wordnet. Using mental lexica as gold standard annotation thus suffers from the same problem as defining the meaning of words by their co-occurrencies or distributions over situations or documents; the derived metrics can be hacked. Also, there's a number of problems with using Wordnets and the like for evaluating word embeddings. The most obvious ones are low coverage and low inter-annotator agreement in such resources. Moreover, as shown by Juergens (2014), some inter-annotator disagreements are not random (errors), but reflect different, linguistically motivated choices. There are different ways to structure word meanings that lead to different semantic networks. Different lexicographic theories suggest different ways to do this. This means that our resources are theoretically biased. After all, while psycholinguistic priming effects and word association norms suggest that semantically similar words are retrieved faster than orthographically similar words, there is to the best of my knowledge no bullet-proof evidence that our brain does not order words alphabetically (or some other obscure way) in the mental lexicon.

Do we have alternatives? Our limited understanding of the brain makes evaluating COGNITIVE LEXICAL SEMANTICS non-trivial – at least if we want to go beyond lexicographic representations of the mental lexicon. If we accept lexicographic resources as approximations of the mental lexicon, we can use these resources for training, as

[2] See Faruqui et al. (2016; Batchkarov et al. (2016; Chiu et al. (2016) for critiques of using word association norms. The problem with word association norms is inadequacy (and statistical power): They conflate several types of similarity, e.g., synonymy and antonymy, and they are culture-specific.

well as evaluation, in the same way as we do evaluation in other supervised learning problems. If we don't, we have to resort to alternatives. Below we consider one, namely direct evaluation against brain imaging (and eye tracking) data.

Denotational semantics At first sight, DENOTATIONAL SEMANTICS seems to assume discrete word representations (sets). Obviously, however, some words have overlapping sets of referents. Can we evaluate our embeddings by how well they predict such overlaps? DENOTATIONAL SEMANTICS, in matrix terms, defines the meaning of a word as its distribution over a set of referents (e.g., its occurrences in Amazon product descriptions). While learning embeddings of words from their distribution over Amazon product descriptions has, to the best of our knowledge, not yet been proposed, this would be easy to do. DENOTATIONAL SEMANTICS is thus very similar to SPRACHSPIEL THEORY from an evaluation point of view; if we fix the set of referents, e.g., Amazon products, evaluation again becomes similar to evaluation in other supervised learning problems.

Brain imaging, anyone? If we accept the premise in the call for papers for this workshop – that down-stream evaluation of word embeddings is impractical and all over the map – we also accept the conclusion that we are interested in embeddings, not only for practical purposes, but as models of cognitive lexical semantics. It seems that this motivates focusing on evaluation procedures such as correlation with word association norms or evaluation against mental lexica. However, lexicographic resources are sparse and theoretically biased, and word association norms are unreliable. What do we do?

If we could measure the semantic processing associated with a word in brain imaging, this would give us a less biased access to the cognitive lexical semantics of words. If we assume such data is available, there are two possible approaches to evaluating word embeddings against such data:

(a) Studying the correlation between distances in word embedding space and EEG/fMRI/etc. space; or, perhaps more robustly, the P@k predicting nearest neighbors EEG/fMRI/etc. using embeddings.

(b) Evaluating the squared error of a regression model trained to associate the input word embeddings with EEG/fMRI/etc.

Note that we have reasons to think such metrics are not entirely inadequate, since we know humans understand words when they read them. fMRI data, for example, may contain a lot of noise and other types of information, but semantic word processing is bound to the contribute to the signal, one way or the other.

At last, a few experiments I present some initial experiments doing both (a) and (b). We evaluate the EW30 and SENNA embeddings (Collobert et al., 2011) against fMRI data from Wehbe et al. (2015), using the token-level statistics derived in Barrett et al. (2016), and eye-tracking data from the Dundee Corpus (Barrett and Søgaard, 2015).

My first experiment is a simple one, merely to show how uncorrelated raw fMRI and eye-tracking data are with state-of-the-art embeddings. I deliberately pick a very simple prediction problem. Specifically, we randomly sample 9 words that are shared between the cognitive gold standard data and the two sets of embeddings we wish to evaluate. For each of the 9 words, I compare nearest neighbors, computing P@1 for both our embedding models.

I convert the fMRI data and the eye-tracking data to vectors of aggregate statistics following the suggestions in Barrett and Søgaard (2015) and Barrett et al. (2016). Table 1 presents the nearest neighbors (out of the 9 randomly selected words) in the gold data, as well as the two word embeddings. The P@1 for both embeddings is 2/9. If I increase the size of the candidate set to 50, and do three random runs, scores drop to 4% and 3.3%, respectively. For comparison, the embeddings agree on the nearest neighbors in 9, 10, and 10 words across three random runs. On the other hand, this is expected, since the embedding algorithms have obvious similarities, while the brain imaging data is entirely independent of the embeddings. If I run the same experiment on the gaze data, using a candidate set of 50 random words, scores are even lower (0–1/50). The P@1 agreements between the fMRI data and the eye-tracking recordings across three runs are also very low (0, 2, and 2 in 50).

If I look at the nearest neighbors across the full dataset, manually, the picture is also blurred. Sometimes, the brain imaging data has odd nearest neighbors, say *teachers* for *having*, when EW30 had *giving*, for example, which is intuitively much closer. In other cases, the gold stan-

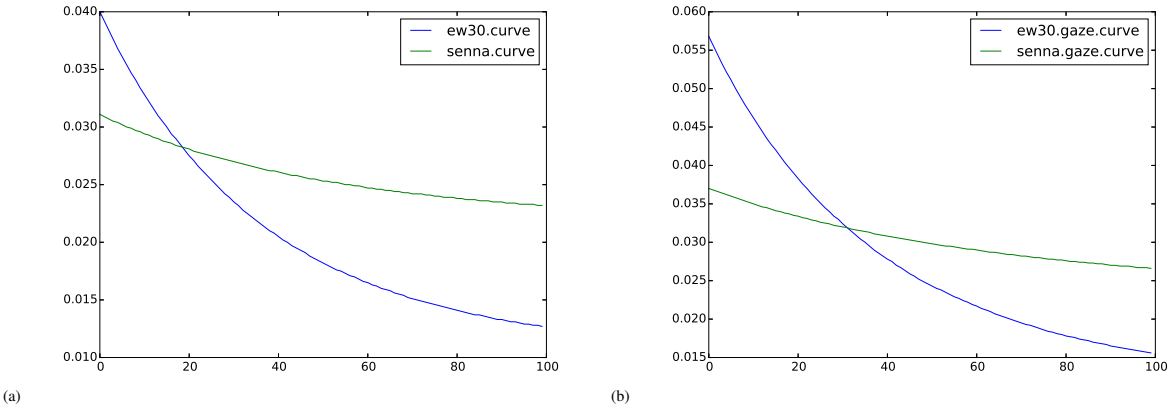

Figure 1: Learning curve fitting state-of-the-art embeddings to token-level fMRI (a) and eye-tracking (b) statistics (x-axis: learning iterations, y-axis: squared mean error)

Target	**Nearest neighbors**		
	SENNA	EW30	GOLD
rolling	nervous	pig	pig
madly	out	nervous	house
rise	hold	hold	anytime
house	hold	pig	anytime
nervous	rolling	rolling	hold
hold	house	rise	managed
managed	hold	out	hold
out	madly	managed	pig
pig	rolling	rolling	rolling

Table 1: Nearest neighbors within a random sample of nine words. We underline the nearest neighbors in SENNA and EW30 embeddings when they agree with the fMRI gold data.

Target	EW30	GOLD	**Okay?**
students	teachers	mistake	No
creep	drift	long	No
peace	death	eat	Maybe
tight	nasty	hold	Maybe
squeak	twisted	broke	Yes
admiring	cursing	stunned	Yes
amazed	delighted	impressed	Yes

Table 2: Examples of nearest neighbors (over full dataset) for EW30 and fMRI embeddings. Manual judgments (**Okay?**) reflect whether the fMRI nearest neighbors made intuitive sense.

dard nearest neighbors are better than state-of-the art, or defendable alternatives. Table 2 lists a few examples, comparing against EW30, and whether the gold standard makes intuitive sense (to me).

However, it is not clear, *a priori*, that the embeddings should correlate perfectly with brain imaging data. The brain may encode these signals in some transformed way. I therefore ran the following experiment:

For words w in a training split, I train a deep neural regression model to reconstruct the fMRI/gaze vector from the input embedding, which I evaluate by its squared error on a held-out test split. All vectors are normalized to the (0,1)-range, leading to squared distances in the (0,2)-range. The training split is the first 100 words in the common vocabulary (of the two embeddings and the gold standard); the test split the next 100 words. Sampling from the common vocabulary is important; comparisons across different vocabularies is a known problem in the word embeddings literature. I use SGD and a hidden layer with 100 dimensions.

I present a learning curve for the first 100 iterations fitting the embeddings to the fMRI data in Figure 1a. Observe that the EW30 embeddings give us a much better fit than the SENNA embeddings. Interestingly, the better fit is achieved with fewer dimensions (30 vs. 50). This suggests that the EW30 embeddings capture more of the differences in the brain imaging data. See the same effect with the eye-tracking data in Figure 1b.

What I am saying ... Under the assumption that downstream evaluation of word embeddings is impractical, I have argued that correlating with

human word processing data is the only valid type of evaluation left on the table. Since brain imaging and eye-tracking data are very noisy signals, correlating distances does not provide sufficient statistical power to compare systems. For that reason I have proposed comparing embeddings by testing how useful they are when trying to predict human processing data. I have presented some preliminary experiments, evaluating state-of-the-art embeddings by how useful they are for predicting brain imaging and eye-tracking data using a deep neural regression model. The test is made available at the website:

`http://cst.dk/anders/fmri-eval/`

where users can upload pairs of embeddings and obtain learning curves such as the ones above. I believe this type of evaluation is the most meaningful task-independent evaluation of word embeddings possible right now. Note that you can also do nearest neighbor queries (and t-SNE visualizations) with the output of such a model.

More advanced theories? Our proposal was in part motivated by a crude simplification of lexical semantics. Of course more advanced theories exist. For example, Marconi (1997) says lexical competence involves both an inferential aspect, i.e., learning a semantic network of synonymy and hyponymy relations, as well as a referential aspect, which is in charge of naming and application. In this framework, a word is defined by its edges in a semantic network *and* its denotation and/or the situations in which it can be used. Technically, however, this is a simple concatenation of the vectors described above. Again, the derived metrics are easily hacked. In other words, if Marconi (1997) is right, evaluation reduces to settling on the definition of the semantic network and of denotation or language games, and finding representative data. From a metrics point of view, any evaluation based on such a theory would be a vicious circle.

Acknowledgments

This work was supported by ERC Starting Grant No. 313695, as well as research grant from the Carlsberg Foundation.

References

Maria Barrett and Anders Søgaard. 2015. Reading behavior predicts syntactic categories. In *CoNLL*.

Maria Barrett, Joachim Bingel, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fmri - with an application to pos induction. In *ACL*.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *RepEval*.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *RepEval*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *ACL*.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *RepEval*.

John Firth. 1957. *Papers in Linguistics 1934-1951*. Oxford University Press.

Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, pages 443–498.

David Juergens. 2014. An analysis of ambiguity in word sense annotations. In *LREC*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225.

Diego Marconi. 1997. *Lexical Competence*. MIT Press.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*.

George Miller and Christiane Fellbaum. 1992. Semantic networks of English. In *Lexical and conceptual semantics*. Blackwell.

Richard Montague. 1973. The proper treatment of quantification in ordinary English. In *Formal philosophy*. Yale University Press.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual nlp. In *ACL*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *EMNLP*.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2015. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, 10(3).

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell Publishing.