Linear discriminative learning

Commentary on R. Harald Baayen et al. (2019). "The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning." In: *Complexity* 2019, p. 4895891

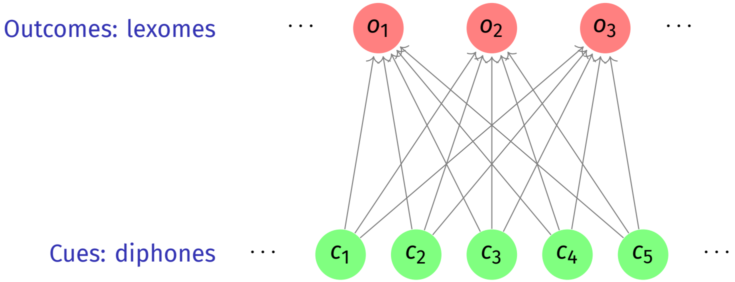
Olivier Bonami

olivier.bonami@u-paris.fr

Université Paris Cité — Lexical Matters — November 2023

The general idea

- In Naive Discriminative Learning models of morphology:
  - Both the cues and the outcomes can be seen as vectors of indicator variables: each cue/outcome is either present (1) or absent (0).
  - *n*-phones as cues capture form similarity, but lexomes as outcomes do not capture similarity of meaning.



- Basic idea of LDL: replace lexomes by distributional vectors.

Morphology as linear algebra I

- Lexical phonological information as a matrix of triphone indicators

$$
\mathbf{C} = \begin{array}{l} \\ \textit{one} \\ \textit{two} \\ \textit{three} \end{array}
\begin{array}{c}
\begin{array}{cccccccc} \#wV & wVn & Vn\# & \#tu & tu\# & \#Tr & Tri & ri\# \end{array} \\
\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}
\end{array}
$$

- Semantic information as a matrix of cooccurrence vectors

$$
\mathbf{S} = \begin{array}{l} \\ \textit{one} \\ \textit{two} \\ \textit{three} \end{array}
\begin{array}{c}
\begin{array}{ccc} \textit{one} & \textit{two} & \textit{three} \end{array} \\
\begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}
\end{array}
$$

Morphology as linear algebra II

- Word comprehension is a matter of mapping correctly from **C** to **S**

$$\begin{array}{cccccccc} \#wV & wVn & Vn\# & \#tu & tu\# & \#Tr & Tri & ri\# \end{array} \qquad \begin{array}{ccc} one & two & three \end{array}$$
$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \overset{\mathbf{F}}{\Longrightarrow} \begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}$$

- Word production is a matter of mapping correctly from **S** to **C** (and then have some algorithm to reconstruct forms from trigrams)

$$\begin{array}{ccc} one & two & three \end{array} \qquad \begin{array}{cccccccc} \#wV & wVn & Vn\# & \#tu & tu\# & \#Tr & Tri & ri\# \end{array}$$
$$\begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix} \overset{\mathbf{G}}{\Longrightarrow} \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

- Linearity assumption: a linear mapping will do.
- This is why the approach is called LDL: the lexicon is modeled by linear transformations between vectors.

Morphology as linear algebra III

- Mathematically, we want to find matrices **F** and **G** such that

$$\mathbf{CF} \approx \mathbf{S} \quad (\text{or } \ \mathbf{SG} \approx \mathbf{C})$$

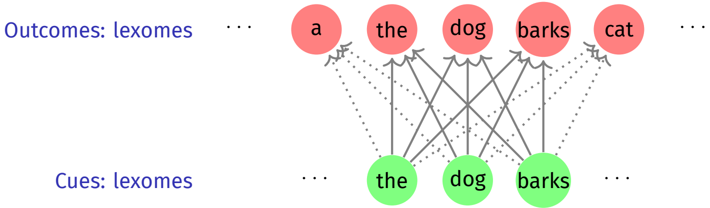
- The Moore-Penrose generalized inverse provides exactly that: a least-squares linear approximation of a function mapping one matrix to another.

$$\mathbf{F} = \mathbf{C}'\mathbf{S} \quad (\text{likewise } \ \mathbf{G} = \mathbf{S}'\mathbf{C})$$

- Note that **F** and **G** represent the outcome of discriminative learning.
  - The authors discuss in passing the fact that such mappings can be learned using the Rescorla-Wagner rule, but they neither demonstrate it mathematically (in this paper) nor discuss psycholinguistic applications involving actual learning.

Deriving semantic vectors

- Instead of using an off-the-shelf algorithm, the authors decided to derive word vectors using the NDL algorithm.
  - The same list of words is used as the cues and outcomes
  - A learning event is a sentence in the corpus.
  - Each word in the sentence counts as a cue to each word in the sentence
    That is, at each sentence
    - The weights from words in the sentence to words in the sentence are upgraded
    - The weights from words in the sentence to words not in the sentence are downgraded



  - The result is a very large $n \times n$ matrix of weights, that ought to be strongly correlated to a matrix of cooccurrence counts.

Semantic vectors: lexomes

- Morphological analysis embedded in the lexomes (derived from TreeTagger + CELEX):
  - Morphologically simplex words contribute a single lexome.
    - *dog* $\rightsquigarrow$ DOG
  - Nonsimplex inflected forms contribute one lexome for the stem + one or more lexome for inflectional categories:
    - *dogs* $\rightsquigarrow$ DOG, PL

    NB: no lexome for SG
  - Nonsimplex derived forms contribute one lexome for the derived lexeme + one lexome for the derivational category (+ lexomes for inflectional categories)
    - *bakers* $\rightsquigarrow$ BAKER, AGENT, PL

    NB: no lexome for base
  - The inventory of inflectional lexomes is clearly motivated by content, e.g. there is a single PAST lexome. The inventory of derivational lexomes is a mixed bag: e.g. separate lexomes for AGENT and INSTRUMENT, but also separate lexomes for ITY and NESS
- Note the absence of structured semantics: sentences *the cats chased the rat*, *the rat chased the cats*, *the cat chased the rats* have identical effects on the vector space.

The semantic vector space

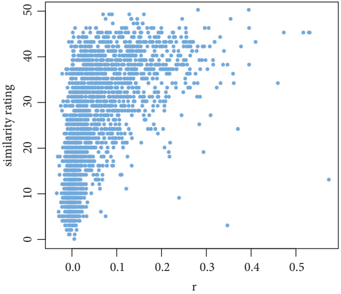
- Vectors derived from the TESA corpus: 750k sentences, 10M tokens, 23,562 lexomes retained for analysis (frequency >8)
- All evaluations rely on the Pearson correlation between vectors as a measure of similarity.
  - In principle, a value between -1 and 1 where:
    1. The absolute value indicates how close we are to a linear relation between the dimensions of the two vectors.
    2. The sign indicates the direction of the slope.
  - In practice, all values are negative $\Rightarrow$ the lower the number, the more similar the vectors.
  - No explanation as to why they use this rather than cosine or Euclidian distance
- Matrix diagonal has highest values, unsurprisingly. For some but not all applications the diagonal values are set to 0.
- All models use a truncated semantic vector matrix, where columns with low variance have been eliminated ($\approx$ 4000 retained columns, varying across models)

Semantic vectors: evaluation I

1. Paired associate learning
   - Psycholinguistic task where participants have to memorize pairs of word and are evaluated on recall of the association.
   - Performance on this task is known to decrease with age.
   - In a linear model, interaction between age and semantic similarity of vectors: the slope of the effect of correlation between the two vectors increases with age.
     - Since correlation is negative, this means that the boost of performance given by semantic similarity in recalling associate decreases with age.
   - Suggests that the vectors do capture something psychologically relevant about similarity between words.

Semantic vectors: evaluation II

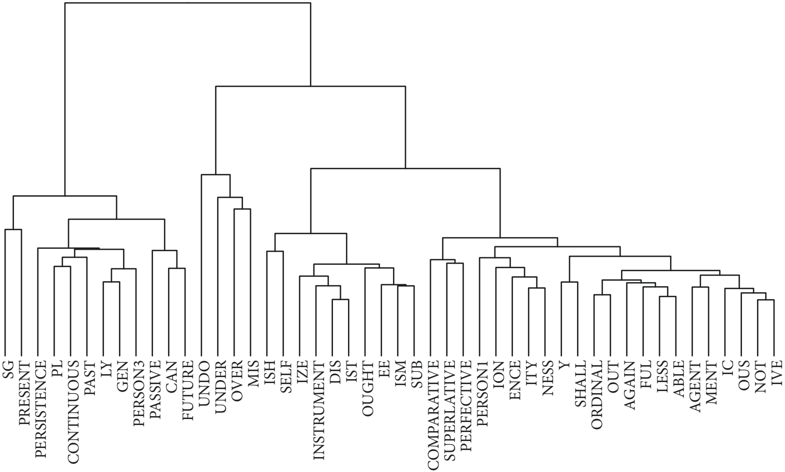
2. Semantic relatedness ratings
   - Interesting relationship between correlation *r* and similarity ratings in the MEN dataset (Bruni, Tran & Baroni 2014).



   - Spearman correlation between MEN scores and *r* between NDL vectors is 0.704.
   - This is slightly better than correlation with LSA scores (0.697). ...but this is much worse than even the 2014 state of the art (Baroni *et al.* 2014), which was at about 0.78

Semantic vectors: evaluation III

3. Correlational structure of morphological vectors



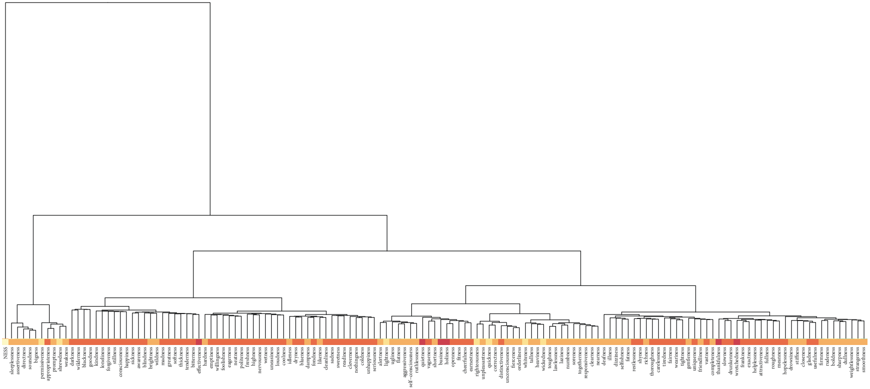
- This feels very close to chance, despite author's optimism.

Semantic vectors: evaluation IV

- The categories shown do not match those described in the paper!
  - Inflectional categories listed in the text:
    COMPARATIVE, SUPERLATIVE, SINGULAR, PLURAL, PAST, PERFECTIVE, CONTINUOUS, PERSISTENCE, PERSON3
  - Derivational categories listed in the text:
    ORDINAL, NOT, UNDO, OTHER, EE, AGENT, INSTRUMENT, IMPAGENT, CAUSER, AGAIN, NESS, ITY, ISM, IST, IC, ABLE, IVE, OUS, IZE, ENCE, FUL, ISH, UNDER, SUB, SELF, OVER, OUT, MIS, DIS
  - Categories present in the heatmap but not described in the text:
    CAN, FUTURE, GEN, ION, LESS, LY, MENT, OUGHT, PASSIVE, PERSON1, PRESENT, SG, SHALL, Y
  - Categories described in the text but not present in the heatmap:
    IMPAGENT, OTHER

Semantic vectors: evaluation V

Interestingly, the category vectors are very different from the vectors for members of the category. E.g. with NESS:

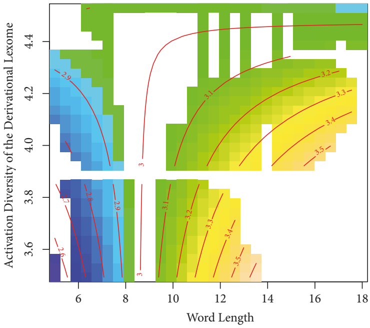


- It is clear why this happens: the weight from NESS to any derivative is downgraded every time a different *-ness* derivative is encountered.
- The authors make a cryptic point implying that this is a good thing.

Semantic vectors: evaluation VI

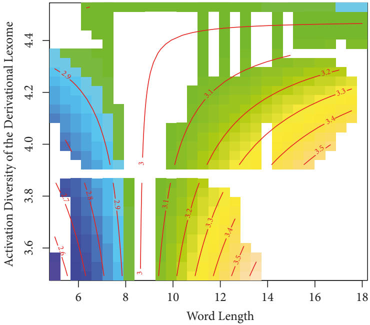
4. Semantic plausibility

- Evaluation against human judgements of semantic plausibility for nonce derivatives from Marelli and Baroni (2015).
- A GAMM showed that word length and activation diversity of the derivational lexome interact in predicting plausibility ratings.
- Remember that activation diversity measures how strongly a cue distriminates among outcomes.
- There are only 6 possible values for activation diversity as there are 6 processes in the dataset.



Semantic vectors: evaluation VII

- Impressionistic examination of correlation between base and derivative suggests reasonable results.
- Evaluation against human judgements of semantic transparency for nonce derivatives from Lazaridou et al. (2016).
- Again, a GAMM showed that word length and activation diversity of the derivational lexome interact in predicting plausibility ratings.
- Note that (at this point) the authors do not have a method to derive a predicted vector for a nonce word, hence the rather coarse-grained evaluation.



- Overall, these vectors are not very impressive.

Comprehension

- Remember: LDL gives us a weight matrix approximating the relationship between form vectors and semantic vectors.
- The authors use this in 4 different ways:
  1. Trigraphs to vectors.
  2. Triphones to vectors.
  3. Trigraphs to triphones to vectors.
  4. Acoustic features of actual speech to vectors.
- Semantic vectors for inflected forms inferred by summing the stem and inflectional lexome vectors.

Comprehension from orthography alone I

- LDL finds **F** such that:

$$\begin{array}{ccccccccccc} \#on & one & ne\# & \#tw & two\# & wo\# & \#th & thr & hre & ree & ee\# \end{array}$$
$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\Downarrow \mathbf{F}$$

$$\begin{array}{ccc} one & two & three \end{array}$$
$$\begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}$$

Comprehension from orthography alone II

- Accuracy: proportion of cases where the closest actual vector to a predicted vector is the correct one.
  - Accuracy on the training set is 59% (compare 27% with NDL)
- Assessment of inflectional productivity: proportion of cases where the predicted vector for an unseen inflected form is closer to the sum of stem and inflectional lexome vectors than to any of the actual vectors.
  - Accuracy is 43% on 553 test items
- The same setup just does not work for derivation: no correlation between predicted vectors and summed stem+derivational category vector.
  - Unsurprising given prior observations on the derivational category vectors.
  - The authors strangely try to argue that this is due to semantic idiosyncrasies in derivation, when they previously established that it is a consequence of their setup.

Comprehension from triphones

- This is the setup I originally described. Find **F** such that:

$$\begin{array}{cccccccc} \#wV & wVn & Vn\# & \#tu & tu\# & \#Tr & Tri & ri\# \end{array}$$
$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$\Downarrow \mathbf{F}$$

$$\begin{array}{ccc} one & two & three \end{array}$$
$$\begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}$$

- Strong boost to accuracy on training data: 78%
  - Compare 59% with trigraphs

Comprehension from speech signal

- We start from a pairing of words with acoustic recordings from the UCLA Library broadcast newsscape.
- From these are derived Frequency Band Summary Features for each token of a word.
- Result: matrix $\mathbf{C}_a$ of 131,673 audio tokens $\times$ 40,639 FBSFs.
- This is put in relation with an expanded semantic matrix where each token of the same type is given an identical row vector.
- Matrix **F** linking the two computed as before, with some complications due to larger matrix size.
- Accuracy evaluated as before: success if actual vector is most highly correlated with predicted vector.
- Result: 34%
  - Compare: 12% with NDL, 6% with Mozilla DeepSpeech.

Production

$$\begin{array}{ccc} \textit{one} & \textit{two} & \textit{three} \\ \end{array}$$
$$\begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix}$$

$$\Downarrow \mathbf{G}$$

$$\begin{array}{cccccccc} \#wV & wVn & Vn\# & \#tu & tu\# & \#Tr & Tri & ri\# \end{array}$$
$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$\Downarrow$$

Candidate forms deduced
from connected sequences
of high activation triphones

Production performance: monolexomic words I

- Evaluation method 1:
  1. From a semantic vector use **G** to obtain a vector of triphone activation weights.
  2. Retain triphones with activation > 0.99.
  3. Construct a directed graph with triphones as vertices and edges between triphones that can be in sequence.

     $\#wV \longrightarrow wVn \longrightarrow Vn\#$

  4. Find the longest simple path (with no repeated triphones) in this graph and deduce a sequence.

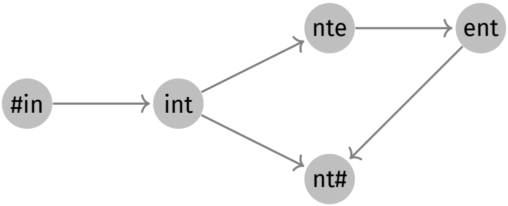
     $\#wV \longrightarrow wVn \longrightarrow Vn\# \quad \Rightarrow \#wVn\#$

  This led to 100% accuracy!
- Problem: this will not work well on novel words, which may contain triphones unseen (or rare) in training, and that will hence never reach the 99% threshold.

Production performance: monolexomic words II

- Evaluation method 2:
  1. Construct a graph with the triphones that are "best supported" by the input vectors (with a complicated definition of "best supported")
  2. Consider all paths from an initial to a final triphone in this graph.

  Accuracy $99,9\%$, all 5 errors being cases where the correct path is not the shortest path whith these triphones.



Production performance: inflected words

- Vectors for inflected words computed by adding stem vector and inflectional function vector.
- Production accuracy of 92%
  - An unknown portion of this is due to inconsistent coding of variation in CELEX.
- 10 fold cross-validation, with training on all stems and 90% of inflected forms.
  - Accuracy 62%.
  - (???) In 3% of cases the correct form is not even a candidate.

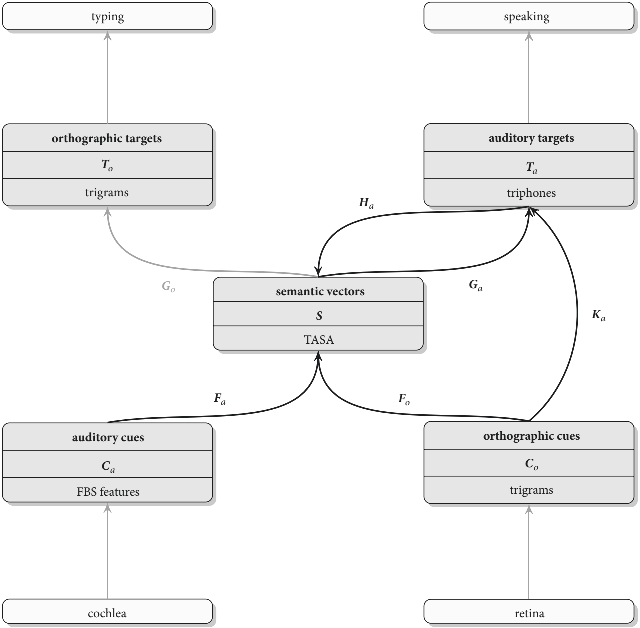
Production performance: derived words

- Starting from the derived word semantic vectors: 99% accuracy.
- Starting from the base vector + derivational category vector: 98.9% accuracy.
- In cross validation, accuracy of 75%

Production: discussion

- The production results are surprisingly good.
  - Small corpus
  - Many prediction errors are due to inconsistencies in CELEX
  - Many prediction errors resemble human speech errors
- Outstanding memorization of existing forms, without any listing of signs.
- The model is in effect dual route: attempts at building a single route network were unsuccessful.

General discussion



General discussion

- Entirely compatible with incremental learning
- Morphology without compositional operations
- Improves on NDL by taking into account semantic similarity
- Scalable: works relatively well with a small corpus
- Not an exemplar-based theory: no explicit representation of exemplars
- Much less complex than deep learning models: no hidden layer.
- Network flexibility: little new data is needed to learn a new pattern

Evaluation

- One super neat new idea: morphology as mapping between vectors.
- Many problems with execution
  - Reporting problems, esp. for the semantic vectors evaluation.
  - Lack of comparison to the relevant state of the art.
  - Incoherence in evaluations (or reporting on how they were chose)
  - Some conceptual problems
    - Notion of 'monolexomic word': how is that Word and Paradigm morphology?
    - Divide between inflection and derivation
    - 'antiprototypical' category vectors
  - So many moving parts...wouldn't we learn a lot more from focusing on just one new idea rather than trying to defend 10 at the same time?