Vector space language models for psycholinguistic analysis

Márton Makrai

Research Institute for Linguistics, Budapest

Corpus resources for quantitative and psycholinguistic analysis
Eger, 2 June 2014

Overview

Introduction

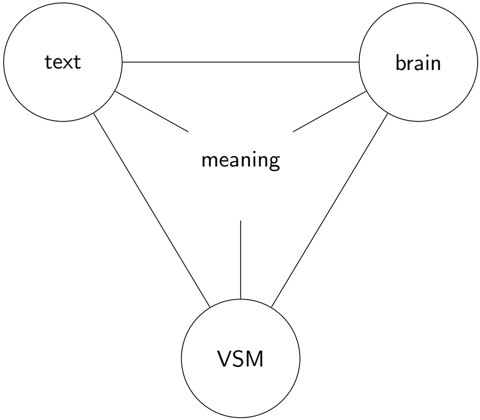
Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

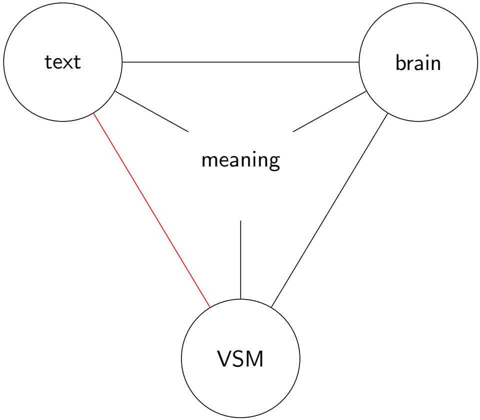
Conclusions

Vector space models (VSMs) for psycholinguistic analysis



Vector space language models (VSM)

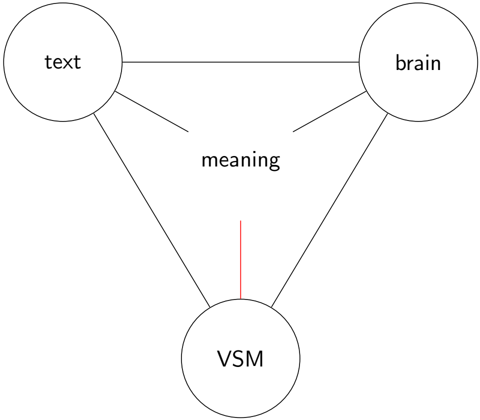
- words are represented by vectors
- dense real-valued vector of some hundred dimensions
- vectors capture different features (syntactic, semantic. . .)
- extend to phrases and short sentences
- compositional morphology


text
brain
meaning
VSM

. . . for natural language processing

- speech recognition (Schwenk, 2007; Dahl et al., 2011)
- many task with the same model (Collobert et al., 2011)
  - language modeling
  - part-of-speech tagging,
  - chunking,
  - named entity recognition,
  - semantic role labeling and
  - syntactic parsing
- searching for images using text (Weston et al., 2010)
- statistical machine translation (Schwenk et al., 2012; Le et al., 2013)
- paraphrase detection (Socher et al., 2011)
- word sense disambiguation (Bordes et al., 2012)
- sentiment analysis (Glorot et al., 2011; Socher et al., 2011b)

Vector space semantics



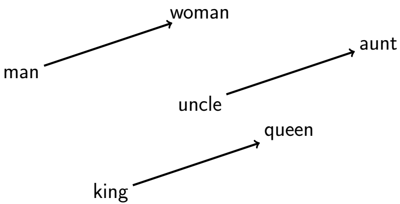
Compositionality in VSM

- compositional representation of words
  - stem + affix

$$\mathbf{v}\,(\text{years}) \approx \mathbf{v}\,(\text{year}) + \mathbf{v}\,(\text{-s})$$

  - lexical decomposition, vector offset method (Mikolov et al., 2013b)

$$\mathbf{v}\,(\text{queen}) - \mathbf{v}\,(\text{woman}) \approx \mathbf{v}\,(\text{king}) - \mathbf{v}\,(\text{man})$$

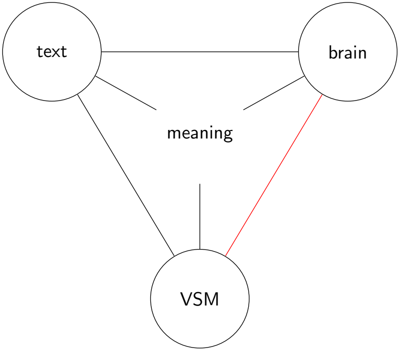


    - at least for frequent words
  - compounds (Wang et al., 2012)

$$\mathbf{v}\,(\text{butterfly}) \stackrel{?}{=} \mathbf{v}\,(\text{butter}) + \mathbf{v}\,(\text{fly})$$

$$\mathbf{v}\,(\text{buttermilk}) \stackrel{?}{=} \mathbf{v}\,(\text{butter}) + \mathbf{v}\,(\text{milk})$$

Psychological reality of VSMs I



- Behavioral tests of word similarity
    - human judgments (synonymy, category membership)
    - feature-norming: participants list the features to words
    - word priming data
- modeling neural activations

Overview

Introduction

Neural modeling

- LSA and brain imaging
- Neural modeling

Morphology, inflections

- The past tense debate
- with VSMs

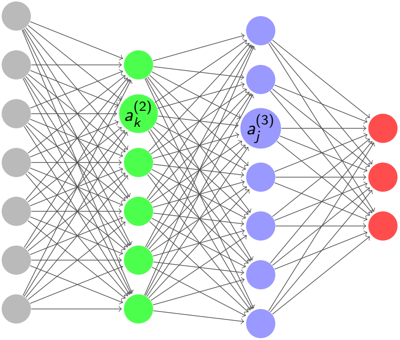
Ambiguity

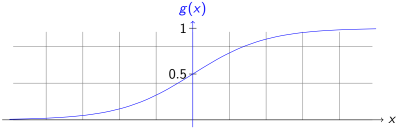
Conclusions

VSM from co-occurrence matrix

- co-occurrence matrices
  - rows are words
  - columns are contexts
    - document (Latent semantic analysis, Huang et al. (2012))
    - words
  - elements: co-occurrence in a small window of words
- dimensionality reduction
- semantic similarity of words $\approx$ cosine similarity of vectors

Neural language model





$$a_j^{(3)} = g\left(\sum_k w_{jk}^{(2)} a_k^{(2)}\right)$$

- machine learning, backpropagation
- Bengio et al. (2003, 2013)
- input layer: 1-of-V
- $w_{jk}^{(1)}$: word embedding (shared by tasks (Collobert et al., 2011))
- output layer
  - nodes represent the probability of each tag

Brain imaging

	what is measures	time res	spatial res
MEG	magnetic field caused by many thousands of neurons firing together	1000 Hz	poor
fMRI	change in blood oxygenation caused by neural activity	0.5–1 Hz	good

Overview

Introduction

Neural modeling

- LSA and brain imaging
- Neural modeling

Morphology, inflections

- The past tense debate with VSMs

Ambiguity

Conclusions

LSA in neurosemantics

- task of neurosemantic decoding (Mitchell et al., 2008): find neural basis images to semantic dimensions
- Pereira et al. (2011)'s system matches the brain images with corresponding articles from Wikipedia
- Palatucci et al. (2009): zero-shot learning of

$$\text{fMRI} \rightarrow \text{word}$$

VSM using brain activations Fyshe et al. (2014)

- incorporate brain activation data recorded while people read words
- complementary strengths of corpus and brain activation data
- predictive power generalizes across brain imaging technologies
- experiments
    - correlation to behavioral data
    - brain $\rightarrow_{\text{lin}}$ word
    - 2 vs 2
    - brain $\rightarrow$ corpus vector (for rare words)
    - data is made available

Overview

Introduction

Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

Conclusions

Ideas from cross-modal mapping

- mapping between vector representations of words and images
- Linguistic motivation: grounding and reference
- searching for images using text (Weston et al., 2010; Socher et al., 2013)
- future: question answering from pictures

Research proposal: mapping between neural language models and brain activation vectors

- choice of the mapping
  - Lazaridou et al. (2014) try 4 learning algorithms
    - Linear Regression
    - Canonical Correlation Analysis (CCA)
    - singular value decomposition
    - neural network
  - neural network: the hidden layer as a cross-modal categorization
  - linear regression
    - non-linearity is already present in embeddings
    - successfully used for translation by Mikolov et al. (2013a)

Overview

Introduction

Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

Conclusions

Overview

Introduction

Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

Conclusions

An eternal topic in distributed language modeling

The past tense debate

- characterization of implicit linguistic knowledge
  - rules
    - Pinker (1984)
    - language acquisition is rule induction
    - innate linguistic universals
    - duality of regular and irregular forms
  - parallel distributed processing (Rumelhart and McClelland, 1986)

The past tense debate

Acquisition of the past tenses of English verbs

- stages
  1. only a small number of verbs in the past tense (high-frequency, mostly irregular)
  2. much more verbs in the past, the majority is regular
     - wug test
     - over-regularization (*comed/camed*)
  3. regular and irregular forms coexist
     - over-regularizations remain
- model

$$\text{phon pres} \rightarrow \text{feat pres} \xrightarrow{\text{neural network}} \text{feat past} \rightarrow \text{phon past}$$

  - feat: feature trigrams
- simulates
  - productivity
  - stages (with transition)
  - etc. (e.g. change of the ratio of $\textsc{past} + \textit{ed}$: $\textsc{pres} + \textit{ed}$ over-regularizations)

Phonological complexity (PC)

- regular past tense verbs involve greater phonological processing
- evidenced by early neuroimaging studies
- regular verbs are phonologically more complex than irregulars
- Oh et al. (2011) experiment manipulating regularity and PC
- main effect of both

Overview

Introduction

Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

Conclusions

Derivational morphology in a VSM

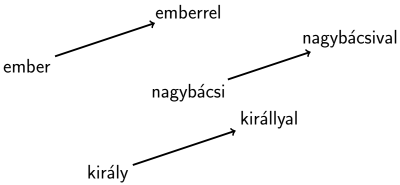
$$\mathbf{v}\,(\text{years}) \approx \mathbf{v}\,(\text{year}) \circ \mathbf{v}\,(\text{-s})$$

- first real-scale system: Lazaridou et al. (2013)
- compositional methods
- building bound morpheme vectors
- morphological analysis given
- experiments:
  - approximating high-quality corpus-extracted vectors
  - comparing the quality of corpus-extracted and compositionally generated words
- future work: composition and morphological induction jointly

Proposal: vector offset analysis of rich inflections

- vector offset method (Mikolov et al., 2013b)

$$\mathbf{v}\,(\text{years}) \approx \mathbf{v}\,(\text{year}) + \mathbf{v}\,(\text{-s})$$



- problems:
  - embedding for a language with a rich morphology
    - Mikolov et al. (2009): speech recognition of Czech lectures
  - inflected words are rare

Overview

Introduction

Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

Conclusions

Ambiguity

- psycholinguistics
  - ambiguous words in different contexts
  - the time course of ambiguity resolution
- VSM: Multiple Word Prototypes (Huang et al., 2012)

Center Word	Nearest Neighbors
bank_1	corporation, insurance, company
bank_2	shore, coast, direction
star_1	movie, film, radio
star_2	galaxy, planet, moon
cell_1	telephone, smart, phone
cell_2	pathology, molecular, physiology
left_1	close, leave, live
left_2	top, round, right

Overview

Introduction

Neural modeling
- LSA and brain imaging
- Neural modeling

Morphology, inflections
- The past tense debate
- with VSMs

Ambiguity

Conclusions

Acknowledgements

- advisor: András Kornai
- Francisco Pereira and Péter Siptár for literature

References I

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. URL http://www.jmlr.org/papers/v3/bengio03a.html.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. PAMI*, 35(8): 1798–1828, 2013. URL http://arxiv.org/abs/1206.5538.

A. Bordes, X. Glorot, J. Weston, and Y. Bengio. Joint learning of words and meaning representations for open-text semantic parsing. In *AISTATS*, 2012.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 2011.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. Large vocabulary continuous speech recognition with context-dependent dbn-hmms. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4688–4691. IEEE, 2011.

References II

Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. Interpretable semantic vectors from a joint model of brain- and text-based meaning. In *ACL14*, 2014.

X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 873–882, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Compositionally derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526, 2013. URL `http://aclweb.org/anthology/P/P13/P13-1149.pdf`.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL-2014*, 2014.

References III

H.-S. Le, I. Oparin, A. Allauzen, J.-L. Gauvin, and F. Yvon. Structured output layer neural network language models for speech recognition. In *IEEE Trans. Audio, Speech & Language Processing*, 2013.

Márton Makrai, Dávid Márk Nemeskey, and András Kornai. Applicative structure in vector space models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 59–63, Sofia, Bulgaria, August 2013. ACL. URL http://www.aclweb.org/anthology/W13-3207.

Tomas Mikolov, Jiri Kopecky, Lukas Burget, Ondrej Glembek, and JH Cernocky. Neural network based language models for highly inflective languages. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4725–4728. IEEE, 2009.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia, 2013b. Association for Computational Linguistics.

T. M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880): 1191–1195, 2008.

Tomasina M. Oh, Keith Liming Tan, Philina Ng, Yeh Ing Berne, and Steven Graham. The past tense debate: Is phonological complexity the key to the puzzle? *NeuroImage*, 57(1):271–280, 2011. URL `http://dx.doi.org/10.1016/j.neuroimage.2011.04.008`.

Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, volume 22, pages 1410–1418, 2009.

References V

Francisco Pereira, Greg Detre, and Matthew Botvinick. Generating text from functional brain images. In *Frontiers in Human Neuroscience*, volume 5, 2011. doi: 10.3389/fnhum.2011.00072. URL http://www.frontiersin.org/human_neuroscience/10.3389/fnhum.2011.00072/abstract.

Steven Pinker. *Language Learnability and Language Development*. Harvard University Press, 1984.

D. Rumelhart and J. McClelland. On learning the past tenses of English verbs. In D. Rumelhart and J. McClelland, editors, *Parallel distributed processing: Explorations in the microstructure of cognition*. Branford, Cambridge MA, 1986.

Holger Schwenk. Continuous space language models. *Comput. Speech Lang.*, 21(3):492–518, July 2007. ISSN 0885-2308. doi: 10.1016/j.csl.2006.09.003. URL http://dx.doi.org/10.1016/j.csl.2006.09.003.

References VI

Holger Schwenk, Anthony Rousseau, and Mohammed Attik. Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 11–19. Association for Computational Linguistics, 2012.

R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP*, 2011b.

R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *International Conference on Learning Representations (ICLR)*, 2013.

Richard Socher, Eric H Huang, Jeffrey Pennington, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809, 2011.

References VII

Hsueh-Cheng Wang, Yi-Min Tien, Li-Chuan Hsu, and Marc Pomplun. Estimating semantic transparency of constituents of english compounds and two-character chinese words using latent semantic analysis. In *Proceedings of CogSci*, page 2499– 2504, 2012.

J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. In *Machine Learning*, volume 81, pages 21–35, 2010.

Connection to GOF-KR

In Makrai et al. (2013) we experiment with a VSM computed from formal definitions of words in a defining vocabulary

- definition matrix:
  - rows and columns correspond to words

$$D_{ij} = \begin{cases} 1 & \text{if } w_j \text{ appears in } w_j \\ 0 & \text{otherwise} \end{cases}$$

- dimensionality reduction from $|V|$ to 50 or 100