Embeddings 2.0: The Lexicon as Memory
MIC3 Working Group
http://www-csli.stanford.edu/MIC3

Hinrich Schütze, email: inquiries@cislmu.org

September 9, 2017

1 Prioritized list of how you can help

1. Read the tentative schedule. Feedback appreciated!
2. Read the Introduction. Feedback appreciated!
3. Look at the questions.
   Feedback and additional questions appreciated!
4. Look at the reading list. Feedback appreciated!

2 Schedule (tentative)

2.1 Tuesday morning

Brief introductions (30 seconds / person)
Presentations I

- Hinrich Schütze (introduction to working group, basically this document, 15+5)
- Jay McClelland (45+15)
- Maja Rudolph: Probabilistic modeling perspective (30+5)
  (Rudolph et al., 2016)

Discussion

2.2 Tuesday afternoon

Presentations II

- Volker Tresp: Tensor memories (45+10)
- Rami Al-Rfou, Marc Pickett: Application perspective (30+5)
  (Pickett et al., 2016; Henderson et al., 2017)
- Andrew Lampinen: One-short learning (15+5)

Discussion

(crazy idea I'm unsure about: use the whiteboard to design a computational memory that is moderately faithful to what we know about human memory; this conversation would establish common ground and everybody would realize which important areas they need to read up on)

2.3 Wednesday morning

Presentations III

- Katrin Erk: One-shot learning (30+10)
  (Wang et al., 2017)
- Felix Hill: Embeddings (30+10)
  (Hermann et al., 2017)
- More talks by MIC3 attendees

Discussion

Model design

WG report to plenary session

3 Introduction

In NLP, we need knowledge about words. This knowledge today often comes from embeddings learned by word2vec, FastText, GloVe and similar methods. I will call these representations "embeddings 1.0" in this introduction, to distinguish them from "embeddings 2.0" whose development the goal of this WG is.

Embeddings 1.0 have many limitations. Three important ones are: (i) embeddings 1.0 do not capture important information that is contained in traditional print (e.g., OED) and computational (e.g., LFG) lexicons; (ii) embeddings 1.0 do not support learning from a single occurrence (i.e., one-shot learning) or from a few occurrences; (iii) embeddings 1.0 do not support gradual adaptation without catastrophic forgetting. (An example for limitation (iii): Assume I have an embedding for the vegetable sense of "squash". I start reading a book about the sport "squash". My embedding of "squash" should adapt to represent the sports sense as I read the book without erasing the vegetable sense.)

Our (simplifying) working hypothesis is that the *lexicon is a form of memory.* Computational modeling of memory has produced much interesting research in *deep learning* that we will review and attempt to exploit.

However, the focus will be to look at *computational models of memory from cognitive science* and to ask: are there architectures, algorithms, ideas from cognitive science that can inspire a new generation of more powerful embeddings or, equivalently, a new generation of computational lexicon that is not hand-coded, but learned from corpora?

Two computational models of memory that are particularly promising are the *complementary learning system (CLS)* and *REMERGE* of McClelland et al.

Four aspects of these models can inform the design of new embedding frameworks.

- CLS is a model of two complementary learning regimes, a slow one and a fast one. This is promising for learning *embeddings of rare words* and for *one-shot learning of embeddings.*

- Current embedding models in NLP use primitive representations of context, mostly cooccurrences. In contrast, human learning of word meaning is based on *a temporal sequence of rich context representations* (anchored in the hippocampus) that are generalized over a period of days and weeks. Replay in sleep and simulations is thought to be important for generalization. Are there elements of our understanding of this aspect of human cognition that we can use for embedding learning models? Content-addressable / key-value memories (Grave et al., 2016; Kawakami et al., 2017) are an interesting type of architecture in this regard that has been proposed in deep learning.

- In CLS, neocortex representations are distributed, overlapping and extract latent semantic structure. These are all properties of word embeddings as well. The difference is that neocortex representations gradually integrate new episodes, i.e., they change. This seems an important property that more powerful lexicon representations also should have.[1]

- CLS supports both *pattern separation* (in the hippocampus) and *pattern similarity* (mainly in the neocortex). Embeddings only support pattern similarity; e.g., if words w_1 and w_2 have the same associational signature, then their embeddings are identical. Pattern separation can keep the representations of w_1 and w_2 separate. Beyond the lexicon, pattern separation can keep two senators separate or two twins separate that are the same in most respects, but need to be assigned different values for a few attributes (year they are up for reelection, spouse). Pattern separation could add a *symbolic component* (or *indices*) to embedding learning models that could group combinations of recurring high-level features together and replace Euclidean space with a more structured and more powerful representational medium.

Potential "Bayesian" connection (i.e., connection to the topic of WG Grounding&Distribution): selection, weighting, attention are crucial for CLS. (i) Selection/weighting of sensory inputs. Which sensory inputs are stored (or linked to) in the hippocampus? (ii) Selection/weighting of memories. Which hippocampus memories are replayed for and transferred to the neocortex? (iii) Selection/weighting of features/memory components. When are two memories/sensory inputs treated as related, when as requiring pattern separation?

4 Questions

4.1 Regarding a new cognitively inspired embedding learning model

Figure 1 suggests that the brain is modularized and has k separate embedding spaces. What are these spaces? What is k?

Modeling similarity in Figure 1's six "modalities" (valance, action, name, motion, color, form) seems straightforward, at least conceptually. But what do representations in the

[1] Learning in NLP and in computational cognitive science models often goes sequentially through the training set in several epochs. This results in interleaving, which is important for learning. But it's hard to argue that these commonly used learning regimes are sensitive to and a good model of the effects of the temporal ordering of experience.

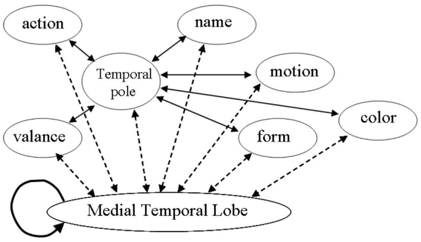

Figure 1: From McClelland (2010): one model of how parts of a memory (or more general parts of any whole?) are bound together

temporal pole look like and what kind of similarity structure do they have? Are representations in the temporal pole sparse or non-sparse?

There is a clear difference between an episodic/temporal index and a entity/concept index in the hippocampus (only the former exists). Is there a difference between the temporal pole constructs that bind the elements of episodes together and temporal pole constructs that bind the elements of entities/concepts together or is there no difference?

Most models in machine learning that are called "memory" models are "drawer" models: a key is used to retrieve a piece of information, for example the word "cat" is used to retrieve the lexical entry for "cat". The McClelland type memory models are not like this. How (not) to reconcile?

Why is content addressable memory efficient and scalable in humans, but not in deep learning?

4.2 Other questions

Various cognitive theories have "binding constructs", constructs that bind different parts of an episodic memory together: hippocampal pattern separation, conjunctive units, indices, integrative representations, concept cells (Quiroga, 2012), convergence zones (Damasio, 1989), the anterior temporal cortex or temporal pole (Figure 1). Are any (or several) of these good models for how different parts of a lexicon entry can be bound together?

What happens with ambiguous words?

Words can have small or big lexical entries. Should they have corresponding small or big embeddings? What type of model would support that?

Ways humans do one-shot learning of word meaning: analogy, lexicon definition, rich event representation (i.e., wampimuk-style, Lazaridou et al. (2014)). How to design models to do this?

5 Readings

5.1 Cognitive science

Kumaran et al. (2016). Relevance: *Updated presentation of complementary learning systems (CLS) theory.* Detailed description of all major aspects of CLS, including slow/fast learning systems, how complementarity supports one-shot-learning without catastrophic forgetting and pattern similarity/separation.

McClelland (2010). Relevance: *This is a high-level presentation of CLS written for a general audience. If you only have time to read one paper, read this one.*

OReilly et al. (2014). Relevance: *Another article that provides an up-to-date description of CLS.*

McClelland et al. (1995). Relevance: *The original CLS paper.*

Kumaran and McClelland (2012). Relevance: *In addition to pattern separation, the hippocampus also is capable of some generalization. The REMERGE model is an extension of CLS that provides an account.* Characteristics of REMERGE: orthogonalized hippocampal episodic codes (instantiated as localist conjunctive units), recurrent processing between feature and conjunctive layers, co-activation during processing of multiple conjunctive units coding for related experiences.

Teyler and Rudy (2007). Relevance: *Clear statement of the indexing / pattern separation function of the hippocampus.* Not a computational model, but contains comprehensive review of supporting literature.

Hinton (1981), **Hinton (1986)**. Relevance: *Introduce the idea of distributed representations of concepts, i.e., the idea of distributional semantics.*

Rabovsky et al. (2017). Relevance: *Recent work on the sentence gestalt (SG) model.* Detailed description and wide-ranging demonstration of empirical adequacy. The SG model can be thought of as one of the first or even the very first model of "embeddings" if we define embeddings as linguistic/language-based distributed representations.

Gaskell and Dumay (2003). Relevance: *Develops experimental methods for measuring the time course of transition of word representations from "fast" "transient" memory to "slow" "permanent" memory.*

Nosofsky (1984). Relevance: *Exemplar theory is a type of simple baseline for CLS models: it has indices, it does slow and fast learning, it can do adaptation without catastrophic forgetting.* (If we wanted to map the core elements of exemplar theory to CLS, would the indices correspond to hippocampus indices and the actual exemplars to neocortex? This is confusing because the REMERGE paper only talks about the relationship between hippocampus and exemplar theory.)

Cer and OReilly (2006). Relevance: *Discusses different ways of binding in the brain: pattern separation in the hippocampus, working-memory binding in prefrontal cortex, "coarse-coded distributed representations" in posterior cortex.*

5.2 Machine learning and deep learning

Hochreiter and Schmidhuber (1997). Relevance: *The short-term memory of long-short-term memory networks is the most common form of memory used in deep learning.* Not theoretically more powerful than recurrent neural networks (Elman, 1990), but much easier to train to keep important information in short-term memory for a large number of time steps. (But large number of time steps is still on the scale of human short-term memory?)

Graves et al. (2016). Relevance: *Describes a type of memory architecture commonly used in deep learning: a controller navigates over an array of memory cells, each containing a vector.* Only working memory, so definitely not a CLS architecture? Focus on "inferior frontal cortex" capabilities?

Blundell et al. (2016). Relevance: *Explicit model of a complementary learning system. Slow learning is modeled as replay of highly rewarding episodes.*

Rudolph et al. (2016). Relevance: *The hierarchical Bayes world meets the embedding world.*

Pickett et al. (2016). Relevance: *The authors are also inspired by CLS, but focus on semantic memory, a type of memory they define as separated from the weights of the neural architecture and stored as "program vectors" in a memory module.*

Chen et al. (2017). Relevance: *It is currently not possible to train large key-value memories, so alternatives are used that are not really desirable, except they are very efficient. Here: a search engine.*

Tresp and Ma (2017), Tresp et al. (2017), Baier et al. (2017). Relevance: *The paper links the knowledge base and semantic web literature to the cognitive science literature.* The authors propose a tensor decomposition formalization of memory. Semantic memory is a triple store. Episodic memory is a quadruple store where each quadruple consists of a triple that was extended with a time/place index. Triples ("Dogs chase cats") are generalizations of quadruples ("Yesterday, my dog chased a cat"). No direct model of the hippocampus, but "index" is a central concept of the theory.

Al-Rfou et al. (2016). Relevance: *This paper presents a good case study that shows that, from an engineering perspective, independent of cognitive motivation, we need episodic memory.* Example: To answer the question "when was reddit user X born", your representation of reddit user X may need to contain a memory of the single time that reddit user X revealed their year of birth.

5.3 Natural language processing (NLP)

Collobert et al. (2011). Relevance: *Highly influential in NLP: not the first to use embeddings and deep learning, but the first comprehensive paper with strong experimental results.* Recall that we interpret embeddings here as a primitive type of lexicon: a word's embedding is its longterm memory representation.

Bahdanau et al. (2014). Relevance: *Apart from embeddings, the other type of memory that is widely used in NLP is "attention". This refers to the ability of keeping the entire input in (working) memory and, at each point in processing, computing a weighted sum over working memory where the weights reflect importance to current context.* This paper introduced this idea. Attention has been extremely successful for machine translation and many other NLP applications.

Kawakami et al. (2017). Relevance: *Good example of effective use of a very small key-value cache (size 100).*

Wang et al. (2017). Relevance: *Models are proposed that, based on properties of the unknown word in a context, can do one-shot learning.*

Hermann et al. (2017). Relevance: *Not really a memory paper, but is relevant because it does gradual learning experience by experience and because, like humans, it does not use cooccurrence, but instead "real" sensory input.*

5.4 Lexicography

Hanks (2015). Relevance: *Probably the best work on contemporary lexicography, with lots of implications for computational and cognitive theories of the lexicon.*

5.5 Clinical research

Vargha-Khadem et al. (1997). Relevance: *Word learning may be possible without hippocampus. It is however hypothesized that perirhinal and entorhinal cortices are needed for word learning.*

Glisky et al. (1986). Relevance: *Learning of a new word is possible without hippocampus if existing knowledge structures can be exploited (my summary).* For example, the ordinary meanings of "run" and "save" can be extended to their computer terminology meanings.

Gabrieli et al. (1988). Relevance: *Was HM able to learn new words?* This article does not provide strong evidence either way.

References

Al-Rfou, R., Pickett, M., Snaider, J., Sung, Y., Strope, B., and Kurzweil, R. (2016). Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baier, S., Ma, Y., and Tresp, V. (2017). Improving visual relationship detection using semantic modeling of scene descriptions. In *ISWC*.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D., and Hassabis, D. (2016). Model-free episodic control. *CoRR*, abs/1606.04460.

Cer, D. M. and OReilly, R. C. (2006). Neural mechanisms of binding in the hippocampus and neocortex: Insights from computational models. In H. Zimmer, A. M. . U. L., editor, *Handbook of binding and memory: Perspectives from cognitive neuroscience*. Oxford.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1:123–132.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Gabrieli, J. D. E., Cohen, N. J., and Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain and Cognition*, 7:157–177.

Gaskell, M. and Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, 89(2):105–132.

Glisky, E. L., Schacter, D. L., and Tulving, E. (1986). Learning and retention of computer-related vocabulary in memory-impaired patients: Method of vanishing cues. *Journal of Clinical and Experimental Neuropsychology*, 8(3):292–312.

Grave, E., Joulin, A., and Usunier, N. (2016). Improving neural language models with a continuous cache. *CoRR*, abs/1612.04426.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwiska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., and Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538:471–476.

Hanks, P. (2015). *Lexical Analysis: Norms and Exploitations.* MIT Press.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y., Lukács, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis, D., and Blunsom, P. (2017). Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In Hinton, G. E. and Anderson, J. A., editors, *Parallel Models of Associative Memory*, pages 161–187. Erlbaum.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Annual Conference of the Cognitive Science Society.*

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Kawakami, K., Dyer, C., and Blunsom, P. (2017). Learning to create and reuse words in open-vocabulary neural language modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1492–1502, Vancouver, Canada. Association for Computational Linguistics.

Kumaran, D., Hassabis, D., and McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20:512–534.

Kumaran, D. and McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, 119(3):573–616.

Lazaridou, A., Bruni, E., and Baroni, M. (2014). Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1403–1414.

McClelland, J. L. (2010). Memory as a constructive process. In Nalbantian, S., Matthews, P. M., and McClelland, J. L., editors, *The Memory Process.* MIT Press.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *J Exp Psychol Learn Mem Cogn*, 10(1):104–114.

OReilly, R. C., Bhattacharyya, R., Howard, M. D., and Ketz, N. (2014). Complementary learning systems. *Cognitive Science*, 38(6):1229–1248.

Pickett, M., Al-Rfou, R., Shao, L., and Tar, C. (2016). A growing long-term episodic & semantic memory. *CoRR*, abs/1610.06402.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nat Rev Neurosci*, 13(8):587–597.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2017). I like coffee with cream and dog? change in an implicit probabilistic representation captures meaning processing in the brain. *bioRxiv*.

Rudolph, M. R., Ruiz, F. J. R., Mandt, S., and Blei, D. M. (2016). Exponential family embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 478–486.

Teyler, T. J. and Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus*, 17(12):1158–1169.

Tresp, V. and Ma, Y. (2017). The tensor memory hypothesis. *arXiv preprint arXiv:1708.02918*.

Tresp, V., Ma, Y., Baier, S., and Yang, Y. (2017). Embedding learning for declarative memories. In *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*, pages 202–216.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Paesschen, W. V., and Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, 277:376–380.

Wang, S., Roller, S., and Erk, K. (2017). Distributional model on a diet: One-shot word learning from text only. *CoRR*, abs/1704.04550.