Computational models of language and processing

Written Research Report Year 1

Ludovica Pannitto

Thu, Sep 5th 2019

34th cycle PhD student - CIMeC - Center for Mind/Brain Sciences

The long-term aim is to provide a **distributional** model of **non-adjacent structures** (i.e., constructions), as they emerge from the **linear linguistic stream** through **general-purpose** statistical learning mechanisms.

Why? Linguistic creativity depends on the ability to re-use existing *chunks* to build up new linguistic instances.
With no boundary between lexical and grammatical level, non-adjacent or partially filled chunks play the biggest part in explaining productivity.

Motivations: Non-Adjacent Dependencies

(1) John **has** play**ed** the piano beautifully.

(2) John **is** play**ing** the piano beatifully.

(3) **The** better**er** John plays the piano, **the more** relaxed we feel.

(4) **John** play**s** the piano beautifully and love**s** to sing.

(5) **Not only** does John play the piano beautifully, **but also** he sings.

(6) John has **studied this piece** for long in order to learn how to **play it** so beautifully.

The emergence of non-adjacent dependencies represents a puzzle, and it is tied to two aspects that cannot be disentangled from linguistic research: the **time-dependent** nature of the linguistic material, and the constraints posed on by **memory and processing**.

Motivations: Segmentation vs. Representation

Computational semantics has often taken segmentation for granted, focusing on **representation**.

Statistical learning (SL) has often ignored the function that chunks play in the utterance, focusing on **segmentation**.

The advent of the newest ANN architectures and multi-modal (e.g., vision) models has shown how addressing the two issues together has a positive impact on the results and on the ability of the models to generalize.

Motivations: Learning

The question about *how do we build and attach meaning representations to linguistic symbols* has for long been central to usage-based models of language acquisition.

In order to be better integrated with the statistical learning and cognitive-based community, we propose to frame the same question in a different formulation:

“How do we identify the linguistic structures that are better suited, or more likely to cue the desired meaning?”

From linearity to hierarchy

The emergence of structure

In a letter-string recall task (Cornish et al. 2017), participants were asked to reproduce a series of 15 string that they had been previously been trained on. The recalled strings were used as inputs for the next participants, in a series of 10 epochs (each involving 10 participants).

Across generations:

- **learnability** of the strings increases: the overall accuracy of the recalled items in terms of normalized edit distance increases, and not at the cost of a collapse of the string sets into shorter sequences
- the **amount of reuse** of chunks significantly departs from randomicity
- **Natural-language like structure** (as compared to a set of strings extracted from the CHILDES corpus) generally emerges

Now or Never bottleneck (Christiansen e Chater 2016)

The *fleeting* nature of memory and the speed of the linguistic input stream creates a bottleneck: the brain must compress and recode linguistic input as rapidly as possible (**Chunk-and-Pass**).

Language acquisition is learning to process rather than inducing a grammar: *acquiring a language requires learning how to create and integrate the right chunks rapidly, before current information is overwritten by new input.*

Moreover, this is not unique to language: e.g., sensory memory is rich in detail but decays rapidly unless it is further processed.

Underlying computational mechanisms

A memory-based perspective (Christiansen e Chater 2016; Altmann 2017) helps in developing a model that takes into account the **relationship between episodic and semantic memory**.

Neurobiologically-inspired models mostly rely on **complementary learning systems** (CLS, McClelland, McNaughton e O'reilly 1995; Schapiro et al. 2017) theory: while the **hippocampal structures** support rapid encoding of different instances, the **neocortex** allows for slower recognition of regularities.

The computational principles by which the learning happens must be able to explain both general tendencies and modality- and stimulus- specific constraints.

Extraction and Integration framework (Thiessen, Kronstein e Hufnagle 2013)

The majority of mechanistic accounts that explain statistical learning focused on sensitivity to conditional relations (i.e., transitional probabilities for word segmentation), ignoring sensitivity to statistical cues (i.e., frequency and variability) that requires integrating information across exemplars.

We can distinguish between two distinct streams (Thiessen 2017), aimed at detecting **conditional** and **distributional** regularities respectively. The former inform a chunk-based memory processes that stores **exemplars**, while the latter are employed to capture **central tendencies** and group elements into categories.

Tools: Distributionalism

Distributional patterns of co-occurrence

The attempt to explain structural properties of language by means of **distributional patterns of co-occurrence** has indeed a long-standing history in linguistic research (Erk 2012; Lenci 2018), with roots in the structuralist distributional analysis (Harris 1954; Braine 1963).

Besides being a quantitative method for semantic analysis, DS could as well be regarded as a **cognitive hypothesis** about the form and origin of semantic representations (Miller e Charles 1991; Lenci 2008), an hypothesis tested also in language acquisition studies (Twomey, Chang e Ambridge 2014; Twomey, Chang e Ambridge 2016).

Distributional methods in statistical learning

"The search for an answer can begin with the cogent assumption that people learn how to use words by observing how words are used." - (Miller e Charles 1991)

Statistical Learning (SL), which had initially focused on word learning (Reber 1967; Saffran, Aslin e Newport 1996), has extended to treating the processing of regularities in sensory input in general, in a more comprehensive theory of information processing (Armstrong, Frost e Christiansen 2017): **experiencers possess the cognitive abilities to take track of distributional patterns**, and this contributes to shaping expectations and behavioral responses.

Tools: Spiking Neural Networks

An issue of plausibility

Artificial Neural Networks (ANNs) and the connectionist paradigm in general have provided a solid framework to implement many of the theories of statistical learning and grammar induction.

ANNs have also been accused of biological implausibility:

- they involve non-local transfer of real-valued errors and weights, while biological neuronal systems assume a kind of firing rate code for transmitting information throughout the brain
- regularities are usually and most effectively extracted through overlapping representations, but non-overlapping item-based representations are equally valuable tools for learning

The Spiking Model

Some of the mentioned drawbacks could be overcome by employing Spiking Neural Networks (SNNs, Maass 1997)[1].

Like ANNs, SNNs are directed graphs made of nodes (*neurons*) and edges (*synapses*).

Interesting features:

- naturally deal with stream-like data over time
- operate using **spikes**, discrete events that take place at points in time, rather than continuous values

[1] a framework is presented in Hazan et al. 2018

The Spiking Neuron

Each biological neuron has a **membrane**, which regulates the production of a spike depending on the received signals.

Using just one variable for modelling the membrane, the state of the neuron at time t is given by its initial state u_0 plus some additional potential due to the received spike stream:

$$u(t) = u_0 + a \int_0^t D(s) \cdot w \cdot \sigma(t - s) ds \tag{1}$$

where a is positive constant, $D(s)$ is a linear filter (e.g., modulates memory loss), w the synaptic weight (excitatory or inibitory) and σ a series of N input spikes, $\sigma(t) = \sum_{i=1}^{N} \delta(t - t_i)$.

A spike is elicited at time t if $u(t) \geq u_{th}$, and the potential is consequently reset to u_0.

Learning algorithms for SNNs: a glimpse

Learning is **local** both with respect to the neighborhood of the synapse and in time, and largely inspired by the basic **Hebbian rule**, (*"cells that fire together wire together"*)

Backpropagation is difficult to apply, both unsupervised and supervised training is possible. The basic idea in the unsupervised case is that the temporal relation between the pre- and post-synaptic spike influences the strength of the connection [2]:

$$\Delta w = \begin{cases} Ae^{\frac{-(|t_{pre} - t_{post}|)}{\tau}} & t_{pre} - t_{post} \leq 0, A > 0 \\ Be^{\frac{-(|t_{pre} - t_{post}|)}{\tau}} & t_{pre} - t_{post} > 0, B < 0 \end{cases} \tag{2}$$

[2] STDP, Spike-timing-dependent plasticity

Learning algorithms for SNNs: a glimpse

Learning is **local** both with respect to the neighborhood of the synapse and in time, and largely inspired by the basic **Hebbian rule**, (*"cells that fire together wire together"*)

Backpropagation is difficult to apply, both unsupervised and supervised training is possible. The basic idea in the unsupervised case is that the temporal relation between the pre- and post-synaptic spike influences the strength of the connection [2]:

$$\Delta w = \begin{cases} Ae^{\frac{-(|t_{pre}-t_{post}|)}{\tau}} & t_{pre}-t_{post} \leq 0, A > 0 \\ Be^{\frac{-(|t_{pre}-t_{post}|)}{\tau}} & t_{pre}-t_{post} > 0, B < 0 \end{cases} \tag{2}$$

[2] STDP, Spike-timing-dependent plasticity

Deep architectures? (Tavanaei et al. 2018)

Only a few tasks have been explored so far, mostly focused on the Modified National Institute of Stan- dards and Technology (MNIST) dataset (LeCun, Cortes e Burges 2010).

Upscaling biologically inspired algorithms such as STDP to more complex architectures still represents a challenge.

Some architectures, as **Liquid State Machines** (LSM, Maass, Natschläger e Markram 2002), are natively equipped with spiking neurons to reproduce the dynamics of cortical circuits.

Few applications to language modeling have been proposed (Costa et al. 2017): although not outperforming LSTMs, **subLSTMs**[3] achieved a comparable level of perplexity in a simple word-prediction task.

[3] LSTM in which the multiplicative gating operations were replaced with substractions

Deep architectures? (Tavanaei et al. 2018)

Only a few tasks have been explored so far, mostly focused on the Modified National Institute of Stan- dards and Technology (MNIST) dataset (LeCun, Cortes e Burges 2010).

Upscaling biologically inspired algorithms such as STDP to more complex architectures still represents a challenge.

Some architectures, as **Liquid State Machines** (LSM, Maass, Natschläger e Markram 2002), are natively equipped with spiking neurons to reproduce the dynamics of cortical circuits.

Few applications to language modeling have been proposed (Costa et al. 2017): although not outperforming LSTMs, **subLSTMs**[3] achieved a comparable level of perplexity in a simple word-prediction task.

[3] LSTM in which the multiplicative gating operations were replaced with substractions

What now?

Some existing models

- R-Grams (Ekgren, Gyllensten e Sahlgren 2018) - based on the *Re-Pair* algorithm (a dictionary-based compression algorithm, Moffat e Larsson 2000), involves the idea that the extraction of abstract chunks or schemas from the input must implement some form of compression
- TRACX2 - (Mareschal e French 2017) - argues that both transitional probabilities learning and chunking can coexist in one system, as it is one single mechanism that underlies sequential learning, Hebbian-style learning. An important aspect is that is that chunks are graded in nature rather than all-or-nothing

R-Grams (Ekgren, Gyllensten e Sahlgren 2018)

Algorithm:

given an initial alphabet of symbols, i.) find the pair *ab* that occurs most frequently in text, ii.) replace all occurrences of *ab* with a new symbol *A*, iii.) add the rule $A \rightarrow ab$ in the grammar, iv.) repeat until no pair occurs more than a defined threshold or the vocabulary size exceeds memory limits.

The **implementation** has a number of drawbacks:

- the complete text is maintained available throughout the whole process
- it's impossible to account for non-adjacent chunks (unless creating a combinatorial explosion)
- it involves a mixture of grammar rules induction and fragments storing: how to perform the parsing phase, if any?

R-Grams (Ekgren, Gyllensten e Sahlgren 2018)

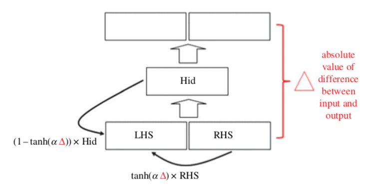
Algorithm:

given an initial alphabet of symbols, i.) find the pair *ab* that occurs most frequently in text, ii.) replace all occurrences of *ab* with a new symbol *A*, iii.) add the rule $A \rightarrow ab$ in the grammar, iv.) repeat until no pair occurs more than a defined threshold or the vocabulary size exceeds memory limits.

The **implementation** has a number of drawbacks:

- the complete text is maintained available throughout the whole process
- it's impossible to account for non-adjacent chunks (unless creating a combinatorial explosion)
- it involves a mixture of grammar rules induction and fragments storing: how to perform the parsing phase, if any?

TRACX2 (Mareschal e French 2017)



$$LHS_{t+1} = (1 - tanh(\alpha\Delta_t) \times Hiddens_t + tanh(\alpha\Delta_t) \times RHS_t$$

Overtime, items that are experienced together become bound to each other and form a chunk. At first it can be a weak, **decomposable** chunk, and later develop into a more self-standing unit.

Both classes of behaviours (i.e., statistical or memory-based) can emerge from a single mechanism: sequence processing emerges from the application of fairly ubiquitous associative mechanisms, coupled with graded top-down re-entrant processing.

Next questions

1. What do (charachter-based) RNNs encode in terms of linguistic structure?
2. Some attempts with SNNs:
   2.1 Can we achieve similar results?
   2.2 What do they encode in terms of linguistic structure?
3. Is there any specific difference in the ability to capture partially filled or non-adjacent constructions?

References

Altmann, Gerry TM (2017). “Abstraction and generalization in statistical learning: implications for the relationship between semantic types and episodic tokens”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711, p. 20160060.

Armstrong, Blair C, Ram Frost e Morten H Christiansen (2017). “The long road of statistical learning research: past, present and future”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711.

Braine, MDS (1963). “The ontogeny of English phrase structure”. In: *Language* 39, pp. 1–13.

Christiansen, Morten H e Nick Chater (2016). “The Now-or-Never bottleneck: A fundamental constraint on language”. In: *Behavioral and Brain Sciences* 39.

Cornish, Hannah et al. (2017). “Sequence memory constraints give rise to language-like structure through iterated learning”. In: *PloS one* 12.1, e0168532.

Costa, Rui et al. (2017). “Cortical microcircuits as gated-recurrent neural networks”. In: *Advances in neural information processing systems*, pp. 272–283.

Ekgren, Ariel, Amaru Cuba Gyllensten e Magnus Sahlgren (2018). “R-grams: Unsupervised Learning of Semantic Units in Natural Language”. In: *arXiv preprint arXiv:1808.04670*.

Erk, Katrin (2012). “Vector space models of word meaning and phrase meaning: A survey”. In: *Language and Linguistics Compass* 6.10, pp. 635–653.

Harris, Zellig S (1954). “Distributional structure”. In: *Word* 10.2-3, pp. 146–162.

Hazan, Hananel et al. (2018). “BindsNET: A machine learning-oriented spiking neural networks library in Python”. In: *Frontiers in neuroinformatics* 12, p. 89.

LeCun, Yann, Corinna Cortes e CJ Burges (2010). “MNIST handwritten digit database”. In: *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist* 2, p. 18.

Lenci, Alessandro (2008). “Distributional semantics in linguistic and cognitive research”. In: *Italian journal of linguistics* 20.1, pp. 1–31.

— (2018). “Distributional models of word meaning”. In: *Annual review of Linguistics* 4, pp. 151–171.

Maass, Wolfgang (1997). “Networks of spiking neurons: the third generation of neural network models”. In: *Neural networks* 10.9, pp. 1659–1671.

Maass, Wolfgang, Thomas Natschläger e Henry Markram (2002). “Real-time computing without stable states: A new framework for neural computation based on perturbations”. In: *Neural computation* 14.11, pp. 2531–2560.

Mareschal, Denis e Robert M French (2017). “TRACX2: a connectionist autoencoder using graded chunks to model infant visual statistical learning”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711, p. 20160057.

McClelland, James L, Bruce L McNaughton e Randall C O’reilly (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.”. In: *Psychological review* 102.3, p. 419.

Miller, George A e Walter G Charles (1991). “Contextual correlates of semantic similarity”. In: *Language and cognitive processes* 6.1, pp. 1–28.

Moffat, NJ Larsson’_and A e J Larsson (2000). “Offline dictionary-based compression”. In: *Data Compression Conference*, pp. 296–305.

Reber, Arthur S (1967). “Implicit learning of artificial grammars”. In: *Journal of verbal learning and verbal behavior* 6.6, pp. 855–863.

Saffran, Jenny R, Richard N Aslin e Elissa L Newport (1996). “Statistical learning by 8-month-old infants”. In: *Science* 274.5294, pp. 1926–1928.

Schapiro, Anna C et al. (2017). "Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711, p. 20160049.

Tavanaei, Amirhossein et al. (2018). "Deep learning in spiking neural networks". In: *Neural Networks*.

Thiessen, Erik D (2017). "What's statistical about learning? Insights from modelling statistical learning as a set of memory processes". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1711, p. 20160056.

Thiessen, Erik D, Alexandra T Kronstein e Daniel G Hufnagle (2013). "The extraction and integration framework: A two-process account of statistical learning.". In: *Psychological bulletin* 139.4, p. 792.

Twomey, Katherine E, Franklin Chang e Ben Ambridge (2014). "Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition". In: *Cognitive Psychology* 73, pp. 41–71.

— (2016). "Lexical distributional cues, but not situational cues, are readily used to learn abstract locative verb-structure associations". In: *Cognition* 153, pp. 124–139.