Do LSTMs know about Principle C?

Jeff Mitchell[1] (jeff.mitchell@bristol.ac.uk)
Nina Kazanina[1] (nina.kazanina@bristol.ac.uk)
Conor Houghton[2] (conor.houghton@bristol.ac.uk)
Jeff Bowers[1] (jeff.bowers@bristol.ac.uk)
[1]School of Psychological Science, University of Bristol, Priory Road
Bristol, BS8 1TU, United Kingdom
[2]Department of Computer Science, University of Bristol, Merchant Ventures Building
Bristol, BS8 1UB, United Kingdom

Abstract

We investigate whether a recurrent network trained on raw text can learn an important syntactic constraint on coreference. A Long Short-Term Memory (LSTM) network that is sensitive to some other syntactic constraints was tested on psycholinguistic materials from two published experiments on coreference. Whereas the participants were sensitive to the Principle C constraint on coreference the LSTM network was not. Our results suggest that, whether as cognitive models of linguistic processes or as engineering solutions in practical applications, recurrent networks may need to be augmented with additional inductive biases to be able to learn models and representations that fully capture the structures of language underlying comprehension.

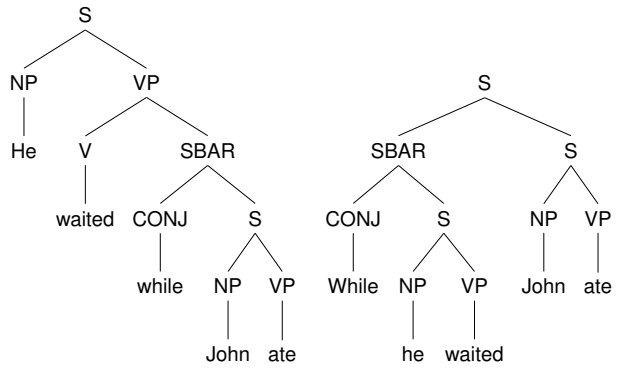
Keywords: Natural Language Processing; Syntax; Recurrent Neural Networks; Psycholinguistics;

Introduction

A recurring theme in connectionist research has been of explaining behaviours which have been claimed to require some innate specifications in terms of learning processes that adapt to the statistical structure of the environment. For example, stage like transitions in development may be explainable in terms of the dynamics of learning in multi-layer networks (Rogers & McClelland, 2004; Saxe, Mcclelland, & Ganguli, 2014).

Language has been a key battleground in these debates, with strong forms of innatism (Chomsky, 1981) and poverty-of-the-stimulus arguments (Pinker, 1979) clashing with approaches that employ simple general purpose learning mechanisms, such as recurrent nets (Elman, 1991). Recently, these nets, and LSTMs in particular, have become the standard tools for building NLP systems and training such architectures on large quantities of raw text has proven to be surprisingly effective way of building representations of natural language data sources (Peters et al., 2018). In addition, interest in the neuroscientific plausibility of these architectures has begun to develop (Costa, Assael, Shillingford, de Freitas, & Vogels, 2017).

The question of their validity as models of linguistic processing has therefore become relevant. Recently, Linzen, Dupoux, and Goldberg (2016) investigated the ability of an LSTM to predict number agreement between subjects and verbs - e.g. *the man eats* but *the men eat* - correctly and found only limited success without explicit supervision. In contrast, Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018) were able to achieve much better performance on the same task, by improving the training procedure.



(a) A tree in which *he* c-commands *John*.

(b) A tree in which *he* does not c-command *John*.

Figure 1: The c-command relation.

Performance at a level comparable to humans on these tasks makes it plausible that an LSTM is capable of identifying the same syntactic dependencies within text that language users exploit in comprehension. However, number agreement is only one of many elements in the complex processes that build sophisticated representations of meaning from raw sequences of words. In particular, understanding the semantics of who did what to whom requires an identification of which entities are being talked about, in which coreference plays a key role.

Here, we explore the extent to which LSTMs are sensitive to these coreference relations using the materials from the psycholinguistic experiments of Kazanina, Lau, Lieberman, Yoshida, and Phillips (2007). In particular, we investigate whether the LSTM trained by Gulordava et al. (2018) is sensitive to Principle C (Chomsky, 1981), a syntactic constraint governing whether a pronoun and noun phrase may corefer. For example, *he* and John cannot corefer in *He waited while John ate*, but can in *While he waited, John ate*.

In experiments 1 & 2, we find that the behaviour of the LSTM is not comparable to the human subjects, and that the LSTM appears to ignore Principle C in anticipating upcoming

Table 1: Sample Items From Experiments 1 & 2.

Condition	Experiment 1	Experiment 2
Principle C/match	Because last semester she$_i$ was taking classes full-time while **Kathryn** was working two jobs to pay the bills, Erica$_i$ felt guilty	It seemed worrisome to him$_i$ that **John** was gaining so much weight, but Matt$_i$ didnt have the nerve to comment on it
Principle C/mismatch	Because last semester she$_i$ was taking classes full-time while **Russell** was working two jobs to pay the bills, Erica$_i$ felt guilty	It seemed worrisome to him$_i$ that **Ruth** was gaining so much weight, but Matt$_i$ didnt have the nerve to comment on it
No constraint/match	Because last semester while she$_i$ was taking classes full-time **Kathryn**$_i$ was working two jobs to pay the bills, Russell never got to see her	It seemed worrisome to his$_i$ family that **John**$_i$ was gaining so much weight, but Ruth thought it was just a result of aging
No constraint/mismatch	Because last semester while she$_i$ was taking classes full-time **Russell** was working two jobs to pay the bills, Erica$_i$ promised to work part-time in the future	It seemed worrisome to his$_i$ family that **Ruth** was gaining so much weight, but Matt$_i$ thought it was just a result of aging

material in the sentence. Experiments 3 & 4 then repeat this analysis on a model trained on data which has been modified to exhibit a stronger signal in respect to Principle C, but the model is still nonetheless insensitive.

In the next two sections, we cover the necessary background relating to coreference and LSTMs.

Coreference and Principle C

Kazanina et al. (2007) investigated the processing of long distance backwards pronominal coreference. They demonstrated that encountering a pronoun leads to an active attempt to identify what the pronoun refers to as soon as possible. If a coreferent was absent from preceding material, then this engendered an expectation that it would arise as soon as possible in future material. This expectation then leads to a so-called gender-mismatch effect, i.e., longer reading times when the main clause subject mismatches in gender with the pronoun (*While she was taking classes full time, Russell was working two jobs to pay the bills.*) as compared to when they match (*While she was taking classes full time, Kathryn was working two jobs to pay the bills.*)

Kazanina et al. (2007) used this effect to show that candidates in illicit structural positions that are subject to Principle C are ruled out of consideration for coreference immediately during incremental processing. Thus, no effect of gender mismatch on reading times would be seen for *She was taking classes full time while Kathryn/Russell was working two jobs to pay the bills*, where the second subject (*Kathryn/Russell*) is ruled out as a candidate coreferent by Principle C.

The particular constraint at work in this example is known as Principle C, and specifies that a pronoun cannot c-command its coreferent. Node N_1 c-commands node N_2 in a syntax tree whenever every node that dominates N_1 also dominates N_2. For example, in Figure 1a the top level S-node is the only node that dominates *he*, and since it also dominates *John*, the pronoun c-commands the name in this example. In contrast, several nodes in Figure 1b dominate *he* that do not also dominate *John*, e.g. SBAR, and so coreference is possible, as the c-command relationship does not hold. This constraint, then, relates to reasonably subtle properties of a sentence's constituency structure, rather than surface features of the word tokens or the linearly intervening material. Additionally, the distance between the pronoun and the noun that are subject to Principle C may in principle be arbitrarily long and they may belong to different clauses. Thus, we might expect this constraint to be more difficult for an LSTM to recognise than the simpler number agreement effects investigated by Gulordava et al. (2018).

Kazanina et al. (2007) ran three experiments on the impact of Principle C on self paced reading times using slightly different constructions in each experiment. We focus here on the first two experiments which considered coreference between pronouns and names, and samples of the materials we analysed are given in Table 1. The intended coreferential dependencies are indicated by subscript indices, and the bold names are the sites where we look for the effect of gender mismatch.

Our interest here is whether an LSTM can learn Principle C from raw text input. This question concerns both the abilities of the LSTM and also the presence in the data of a learnable signal. The c-command structure may be more difficult for the LSTM to identify than the dependency between a verb and its subject, but the statistical trace of Principle C in the training data may also be less clear. Whereas number agreement is compulsory for a verb and its subject, leading to a strong statistical signal in the data, Principle C is a constraint on coreference and only specifies when a pronoun and name must not corefer; in the cases where it does not apply coreference is possible but not required. In this case, we can expect the statistical signal to be weaker due to the presence of cases which could corefer but do not.

Long Short Term Memory Networks

The Long Short Term Memory (LSTM) architecture was proposed by Hochreiter and Schmidhuber (1997) as a model for sequential data to address the problem of vanishing gradients in simple recurrent nets, which had made learning long term dependencies difficult. The main innovation was to use memory cells which maintained a constant memory trace across time and a set of gates to control what flowed into and out of these cells. This approach and its variants (Gers, Schmidhuber, & Cummins, 2000; Cho et al., 2014) is now widely applied to sequential data, and has been used extensively in NLP. Training an LSTM on large quantities of raw text in a language model setting, where the objective is simply to predict a word from its context, has been shown to induce representations that capture much of the relevant linguistic structure, and provide a strong input to downstream supervised tasks (Peters et al., 2018). In terms of modelling human language processing, Goodkind and Bicknell (2018) found that the log probabilities from such a model were effective predictors of human reading times.

Gulordava et al. (2018) trained their LSTM language model on around 3M sentences of English, and demonstrated it was able to predict number agreement between subjects and their verbs. They showed that this grammatical ability persisted even when the sentences were semantically implausible, and that performance of the model on increasingly difficult examples closely mirrored human accuracy. The following experiments investigate whether this human-like performance can also be found in its handling of coreference relations.

Experiments 1 & 2

In this section, we attempt to reproduce Experiments 1 & 2 of Kazanina et al. (2007) using the pre-trained LSTM of Gulordava et al. (2018). We look for the influence on the log probabilities of a gender mismatch between a pronoun and a subsequent name, and attempt to determine whether this response differs between cases where coreference is licit or illicit. In the human experiments, only the cases where coreference was not ruled out by Principle C - i.e. where the pronoun did not c-command the name - resulted in increased reading times associated with a gender mismatch. In the cases where the pronoun did c-command the name - i.e. where coreference is illict under Principle C - reading times were unaffected by a gender mismatch between the pronoun and name.

The signature, then, of a model that is sensitive to the Principle C constraint on coreference will be that it adjusts the likelihood of a dependency between pronoun and name gender depending on whether the former c-commands the latter. In particular, we should expect that agreement in gender is more likely when coreference is possible, i.e. when the pronoun does not c-command the name. In other words, like the human subjects, we expect the model to be more surprised by a mismatch in gender, when it encounters a name that is otherwise a syntactically licit candidate for coreference. Statistically, we analyse this effect in terms of an interaction between gender mismatch and the c-command relation, producing a shift in the logs of the probabilities given to male and female names by the LSTM.

Starting from a list of common names drawn from the register of births in England and Wales (Bush, Powell-Smith, & Freeman, 2018), we found the most frequent 100 male and 100 female names in the training data used by Gulordava et al. (2018). We then ran their LSTM on the materials from Experiments 1 & 2 of Kazanina et al. (2007), obtaining log probabilities for the whole vocabulary at the site of the relevant name in each sentence. We then regressed the log probabilities of the 200 frequent names, on features representing the gender of the name, the gender of the preceding pronoun, whether these genders matched and finally whether the pronoun c-commanded the name.

For the materials from both experiments, the main effects of the first three of these features were very significant ($p < 0.001$), with similar effect sizes in both cases. To summarize, male names are overall more probable than female names, and matching pronoun and name genders are more probable than mismatching. There is also a main effect of pronoun gender which makes names less probable overall following a male pronoun. The direction of the main effect of the c-command feature varied between Experiment 1 and Experiment 2, reflecting the different constructions used in the two sets of materials. For Experiment 1 the regression showed a very significant ($p < 0.001$) increase in the probabilities of names within the c-commanded constructions, but this effect was a less significant ($p < 0.01$) reduction in probabilities for Experiment 2.

In terms of an effect analogous to the longer reading times for gender mismatched names in coreferentially licit positions, we need to look at the interaction of the gender mismatch and c-command features. For Experiment 1, this interaction is significant ($p < 0.01$), but of the wrong sign. In this case, probabilities are lower for gender-matching pronoun-name pairings when coreference is licit than when it is ruled out by Principle C. For Experiment 2, the sign is correct, but the effect is only weakly significant ($p < 0.05$). In both cases, the magnitude of the coefficient for the interaction term is at least 5 times smaller than the main effect of gender mismatch. In other words, the model mainly reacts to a general gender mismatch between pronouns and names, and only weakly and inconsistently takes into account their syntactic relationship.

In a linear regression of the log probabilities against the same set of features on the concatenation of both datasets, the main effects for all features are now very significant ($p < 0.001$). However, the interaction term is now not significant even at the $p < 0.1$ level.

Experiments 3 & 4

Experiments 3 & 4 repeat the analyses described above with an LSTM trained on a modified dataset. Two explanations for the lack of sensitivity to Principle C seen in Experiments 1 & 2 are possible. The first is that the LSTM architecture is not suitable for learning the required structure, while the second is

that the relevant signal is not present in raw text. Perhaps the LSTM fails to adjust its probabilities consistently in response to Principle C simply because there was nothing to be gained from doing so during training.

In fact, we found that the training data contained a weak, though nonetheless significant, interaction between gender mismatch and the c-command structure, leaving open the question of whether the LSTM would have learned Principle C given a stronger signal. Here, we examine whether we can force the LSTM to pay attention to this constraint, by making the relevant signal in the data much more statistically salient.

In particular, we modify a subset of training sentences to always enforce a gender match between pronoun and name when coreference is licit. We find those sentences containing one of the pronouns *he* or *she* followed somewhere further on by one of the 200 names employed in the previous experiments. We then use the AllenNLP tools (Gardner et al., 2017) to obtain constituency parses for these 53K sentences and thus identify those cases where the pronoun c-commands the name. Leaving these cases where coreference is illicit alone, we substitute in a random name from our 100 most frequent male names to the candidate slot following every male pronoun, and a random name from the 100 most frequent female names following every female pronoun. We then retrain the LSTM on the modified 3M sentences training set using the code from Gulordava et al. (2018) and repeat the analyses applied to Experiments 1 & 2.

For both sets of materials, all the main effects are very significant ($p < 0.001$), but the interaction term is not significant even at the $p < 0.1$ level. In other words, making the signature of Principle C more salient in the training data, has not enabled the LSTM to take advantage of this constraint more effectively in making predictions.

Conclusions

Our experimental results indicate that the LSTMs did not learn the Principle C constraint on coreference when trained as language models on large amounts of raw text. This was true both for the original training data and also for the modified version in which the statistical signature of Principle C was made much more salient. In contrast, children as young as 3 do respect this constraint (Lust, Eisele, & Mazuka, 1992), whether acquired innately or by learning from the input. Future work will investigate what additional enhancements are needed to enable neural networks to handle these coreference structures effectively.

Acknowledgments

This research was supported by the European Research Council Grant Generalization in Mind and Machine, ID number 741134. CJH is supported by a James S. McDonnell Foundation Scholar Award in Cognition (JSMF #220020239).

References

Bush, S., Powell-Smith, A., & Freeman, T. (2018, 10 31). Network analysis of the social and demographic influences on name choice within the uk (1838-2016). *PLoS ONE*, *13*(10).

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014* (pp. 1724–1734). Doha, Qatar: ACL.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Costa, R. P., Assael, Y. M., Shillingford, B., de Freitas, N., & Vogels, T. P. (2017). Cortical microcircuits as gated-recurrent neural networks. In *Proceedings of NIPS 2017* (pp. 271–282). USA: Curran Associates Inc.

Elman, J. L. (1991, September). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195–225.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., ... Zettlemoyer, L. S. (2017). AllenNLP: A deep semantic natural language processing platform. *arXiv e-prints*, arXiv:1803.07640.

Gers, F. A., Schmidhuber, J. A., & Cummins, F. A. (2000, October). Learning to forget: Continual prediction with lstm. *Neural Computation*, *12*(10), 2451–2471.

Goodkind, A., & Bicknell, K. (2018, January). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th CMCL workshop* (pp. 10–18). Salt Lake City, Utah: ACL.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018, June). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL 2018* (pp. 1195–1205). New Orleans, Louisiana: ACL.

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

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language*, *56*(3), 384–409.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*, *4*, 521–535.

Lust, B., Eisele, J., & Mazuka, R. (1992). The binding theory module: Evidence from first language acquisition for principle c. *Language*, *68*(2), 333–358.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedins of NAACL 2018.*

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*, 217-283.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA.: MIT Press.

Saxe, A. M., Mcclelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In *In proceedinhs of iclr.*