Relative Clause Semantic Role Labeling Project

Will Howe
whowe1@jhu.edu

Abstract

Across languages, children acquire subject relative clauses earlier and with fewer errors than object relative clauses. The asymmetry surfaces in adults as a processing delay between SRs (subject relatives) and ORs (object relatives). In this project, I investigate whether neural nets exhibit this learning asymmetry. I use SRL (Semantic Role Labeling) as a proxy task for determining if a model has *comprehended* a given relative clause.

1 Introduction

Recent work discussed in (Linzen and Baroni, 2020) and (Kirov and Cotterell, 2018) has shown that neural nets, despite earlier shortcomings as cognitive modeling tools, may be useful as models of human language processing. It is well documented that children face difficulties in acquiring ORs (Object Relative Clauses) from child acquisition studies. These two types of relative clauses can be generated from a single event predicate with two arguments as seen below.

(1) a. The cat that __ ate the mouse. (SR)
b. The mouse that the cat ate __. (OR).

We know the learning asymmetry persists into adulthood from studies on adult parsing (Belletti and Guasti, 2015). Thus, if one believes neural nets exhibit human-like performance on language-domain tasks or ought to, there is a genuine question whether such learning errors and asymmetries will arise in these architectures. Here I will approach this problem from a language acquisition standpoint, while also incorporating tools and analyses from event semantics.

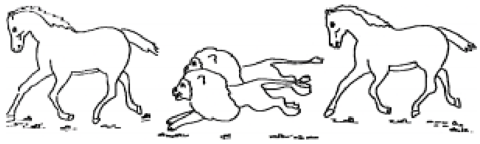
2 Theoretical Background

There are various theoretical analyses for the presence of the SR/OR learning asymmetry. However, the dominant line of inquiry comes from generative grammar.

2.1 Child Acquisition Studies

There is a long line of child acquisition studies demonstrating the SR/OR asymmetry. In a study by (Friedmann and Novogrodsky, 2004), children were read out either an OR or SR sentence and were asked to choose between two pictures:


The four year olds in the study achieved a comprehension rate of 85% on subject relatives, but only comprehended object relatives at chance. The authors claim that the errors are the result of a failure to assign the correct thematic role to the *moved* constituent. Due to problems with the experimental setup in (Friedmann and Novogrodsky, 2004), a new experiment was conducted by (Adani, 2011) where children were asked to point to a referant according to an SR sentence or an OR sentence.



Children were asked, ”Point to the horse the lions are chasing or point to the horse that chases the lions” to generate both possibilities.

2.2 Generative Grammar

The generative grammar analysis of the empirical evidence assumes a raising analysis for both ORs and SRs. Here an embedded matrix clause with a verb predicate and two arguments can be converted into a subject relative if the subject position argument raises into a complementizer phrase headed by a complementizer such as *that, who, what, which, et al*. The corresponding relative clause can be generated by moving the matrix clause object instead of the subject. The acquisition asymmetry then arises from locality constraints on phrasal movement. In fact the child acquisition experiments above form much of the basis of the theory of featural relativized minimality whereby if the grammatical features (gender, case, number, et. al) of the moved referent are *richer* than those of the intervening referent, the sentence will be easier to process.

It is clear then, from the studies in (Adani, 2011) and (Friedmann and Novogrodsky, 2004) that while the SR/OR learnability asymmetry is largely due to syntax, it is sensitive to other grammatical features and is thought to involve thematic role assignment by some. Thus, it makes sense to test neural nets for comprehension of relative clauses which sit squarely at the syntax-semantics interface.

2.3 Tomasello

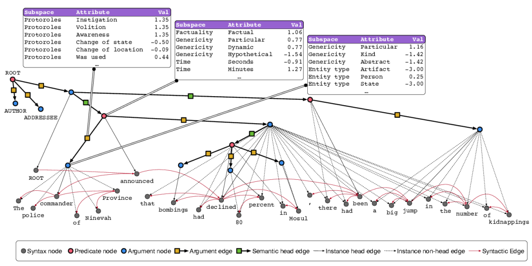
An alternative analysis is put forth by (Diessel and Tomasello, 2005) based in Tomasello's usage based theories of language acquisition. His analysis attempts to explain the core asymmetry by relating various relative clauses (SR, OR, and others) to the frequency of other simple sentences in child directed speech. Because my project is trained on data sources like the English Web Treebank and a dataset called RELPRON, the theories of Tomasello won't factor into my analysis. More specifically, Tomasello considers five types of RCs instead of the two basic types (SR/OR) found in the RELPRON dataset rendering many of his arguments inapplicable here.

2.4 Animacy Effect

As demonstrated in (Goodluck and Tavakolian, 1982), there is an *animacy effect* to consider when dealing with RCs. When a child is presented with the sentence *The lion kisses the duck that hits the pig*, they will sometimes confuse the lion as the argument to the verbal predicate, *hit*, instead of the true subject, the duck.

3 Dataset

I make use of two datasets for assessing LSTM acquisition of relative clauses. First is the Semantic Proto Roles component of the Universal Decompositional Semantics Dataset. Given that the (Friedmann and Novogrodsky, 2004) conception of the SR/OR asymmetry involves thematic role labeling, and since broadly the task is sensitive to animacy, and other features of NP constituents, it makes sense to use a dataset which decomposes thematic roles into feature bundles which can be combined in various combinations and scalar values to recover standard thematic roles.



Since I use Semantic Role Labeling (SRL) as a proxy for probing whether a model as *comprehended* a RC (Relative Clause), the UDS dataset is useful for training a model with a representation of thematic roles as proto-role feature bundles a la (Reisinger et al., 2015).

Secondly, I make use of a dataset called RELPRON which consists of sentences in the following form where OBJ stands for Object Relative and SBJ stands for Subject Relative (Rimell et al., 2016)

```
OBJ garrison_N: organization_N
that government_N station_V

SBJ garrison_N: organization_N
that defend_V castle_N
```

Originally used for research on compositional distributional semantics, the dataset is easily translated into a sentence of the following structure:

```
A garrison is an organization
that a government stations
```

I apply basic text processing and a spelling correction Python library to each sentence to generate the above natural language pattern which is a noun phrase followed by a copula followed by an SR or OR which effectively offers a definition of the original word. Thus, the complement of the copula can be used to generate two predicate-argument pairs, one for the proto-Agent argument and one for the proto-Patient argument. Since the syntax of these RCs largely matches up with the semantics (i.e. the subject position largely matches up with the proto-Agent thematic role) I use this heuristic for labeling my predicate-argument pairs.

4 Procedure

In addition to testing hypotheses about the learnability of RCs by LSTMs, my project also explores the effects of pre-training a model on UDS.

4.1 Pre-Training

I pre-train a model to perform SPRL (Semantic Proto Role Labeling) using UDS SPR (Semantic Proto Roles) dataset built on the English Web Treebank. More specifically, I adopt a similar model to the one detailed in (Rudinger et al., 2018). My modeL uses Glove embeddings (Pennington et al., 2014) of dimension 300 followed by a biLSTM of hidden layer dimension 25. Then the predicate and argument encodings are concatened to form a 100 dimensional vectors which is passed to a linear layer with an output of size 20. Then for each of the 14 proto-role attributes in UDS a scalar value is predicted and then MSE loss is computed for each attributed and summed up. Once this model has been pre-trained on the UDS data, I make several assumptions; firstly that multi-regression prediction of SPR features will force the model to learn a representation of semantic proto-roles. Next, since UDS features can be used to derive the roles of agent and patient, I will assume that the model's internal representation is sufficient for differentiating the canonical roles, Agent and Patient. While I intended to remove all instances of RCs from the UDS data prior to pre-training, this was not feasible for this particular iteration of the project since RCs can occur in a number of forms that can be hard to differentiate and detect in a dataset like the English Web Treebank (EWT). Additionally, the data in the RELPRON dataset is tailored to a very specific linguistic phenomenon which likely does not occur in isolation in UDS which may minimize the initial concern.

4.2 Fine-Tuning

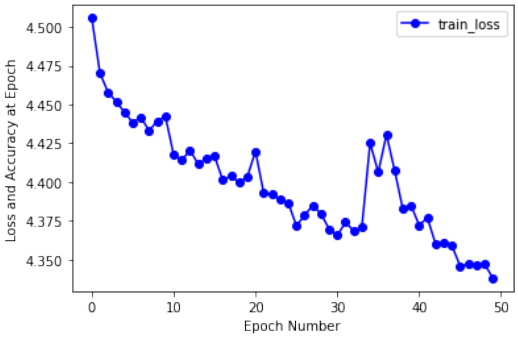
For my fine-tuning phase, I modify the SRPL architecture to predict the categorial label, Agent or Patient by replacing the multi-regression heads with a MLP (multi-layer perception) feed-forward block. Over the course of training, I am able to inspect my loss curves to check for the existence of an SR/OR asymmetry. During this fine-tuning phase I will train strictly on the post-processed RELPRON data. To better understand the effects of a UDS-pretrained model, I will compare the results this model to an out-of-the-box biLSTM model that has not been pre-trained and compare their respective performance.

5 Results

5.1 Pre-training

On the initial pre-training of UDS using the Neural-Davidsonian Semantic Role

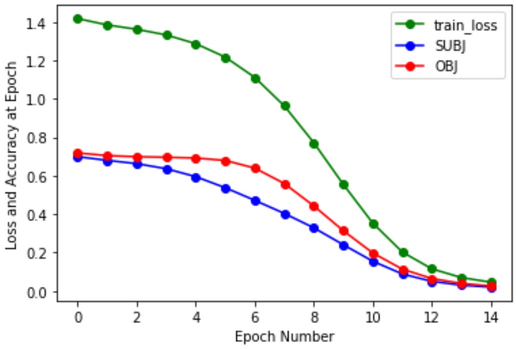
Labeling model, I am unsurprisingly able to train my model to convergence using a small subset of the SPR-UDS data.



Since I train using a batch size of 10 and since the overall MSE training loss is calculated as a sum of the losses for the 14 attributes, the model clearly achieves a low MSE on the subset of the UDS data.

5.2 Bare biLSTM

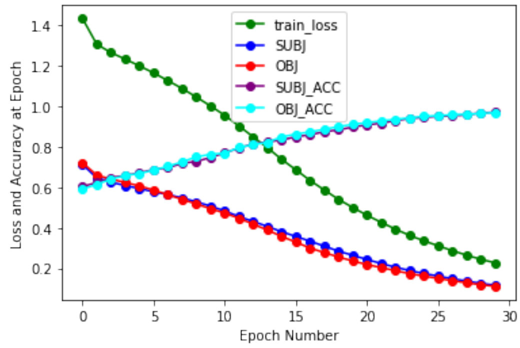
The untrained LSTM model replaced the 14 linear layers from the pre-trained model with a single linear layer for categorial prediction. While at first my loss curves seemed to indicate that the bare biLSTM model was learning as predicted by my hypothesis (exhbiting an SR/OR distinction) upon further inspection, this was not the case.



While the loss for object relatives is consistently higher than that of subject relatives, the model achieves 100% accuracy on both classes or RCs almost immediately. These loss curves might have been able to be taken for evidence of the asymmetry had the accuracy curves also exhibited a similar effect, but this was not the case. Instead I am left to conclude that there is an anomaly in the dataset or in the network architecture that creates this discrepancy. Ultimately though the discrepancy appears visually between the loss curves, it is too small (.05) to draw any conclusions.

5.3 Pre-trained biLSTM

In contrast to the bare biLSTM, the pre-trained LSTM is trained on the multi-regression task of predicting Semantic Proto Role attributes. Consequently, this model has built a representation focused on the semantics of the input almost exclusively.



Whereas the base biLSTM almost immediately reached 100% accuracy on this task (likely by learning a strictly syntactic representation) the UDS pre-trained model had accuracy on both SRs and ORs in the 60% range. This is because the pre-trained model has reached a local optimum on the purely semantic task and will require more training to move from the semantic domain into a more syntactic domain.

6 Analysis

My initial hypothesis on this project was that the bare LSTM model would exhibit a human-like acquisition pattern in its comprehension of RCs. However, its clear that this was not the case insofar as semantic role labeling is a good proxy for measuring *comprehension* of an RC. If the bare LSTM model had performed according to my original hypothesis, I would have imagined that the UDS-pretrained model might ameliorate some of the asymmetry by improving performance by means of a more fine-grained representation of semantics. Since this was not the case, it is hard to make sense of the results of the task on the pre-trained model.

One fruitful path forward might be to analyze qualitatively the examples that model gets wrong. However, after a first pass, I don't find a clear pattern or generalization in errors. For example, at epoch 30 of training the model gets the sentence: *A temple is a building that a pilgrim visits* and *A mother is a woman that a child loves*. In the first example it could be argued that there is a animacy effect due to the word temple, but the model also gets wrong the second sentence in which both referents are animate and whose predicate *love* can reasonably be thought to assign an Agent and Patient role.

7 Future Work

While it is clear that the task of Semantic Role Labeling (SRL) alone is not a good proxy task for determining whether a model has comprehended a RC, I have two ideas for ways to improve the experiment.

- First if it is the case that the non pretrained model did so well due to 1) the LSTM architecture in conjunction with 2) pretrained Glove word embeddings, using word embeddings that are compressed to only include semantic content and no syntactic content will provide a better model of human acquisition (Li and Eisner, 2019).

- Secondly, to make the task harder, a similar dataset could be created which has equal parts sentences whose thematic role assignments match the syntax and those whose do not. For example the RC, *A garrison is an organization that withstands a siege*. Here the siege could be a collective noun standing in for a group of soldiers, taking the Agent role while the garrison would be the Patient. Such a dataset would not be comprehended so easily by the LSTM and might get more closely at the root of the problem. However, it might also be that the model would perform well on syntax-semantics congruent RCs and less well on these incongruent ones. Ultimately, the original hypothesis has been falsified though because children have acquisition problem even with these so called syntax-semantics congruent Object Relatives.

8 Code

I conducted most of work using Google Colab with GPU runtime enabled to enable faster training. My models are trained in Pytorch with the Adam optimizer. Code and data are accesible here:

`https://github.com/ahowe444/RC_SRL`

9 Conclusions

In this project, I posited that neural nets would exhibit a human-like asymmetry in acquisition of relative clauses, but this was not the case. There are certain modifications that could be used to probe the matter further, but ultimately they won't have much bearing on the original hypothesis. I also learned that for this task UDS pre-training was not a useful pre-training strategy though it might be useful for other tasks.

References

Flavia Adani. 2011. Rethinking the acquisition of relative clauses in italian: towards a grammatically based account. *Journal of child language*, 38(1):141.

Adriana Belletti and Maria Teresa Guasti. 2015. *The Acquisition of Italian: Morphosyntax and its interfaces in different modes of acquisition*, volume 57. John Benjamins Publishing Company.

Holger Diessel and Michael Tomasello. 2005. A new look at the acquisition of relative clauses. *Language*, pages 882–906.

Naama Friedmann and Rama Novogrodsky. 2004. The acquisition of relative clause comprehension in hebrew: A study of sli and normal development. *Journal of Child language*, 31(3):661.

Helen Goodluck and Susan Tavakolian. 1982. Competence and processing in children's grammar of relative clauses. *Cognition*, 11(1):1–27.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. *arXiv preprint arXiv:1910.00163*.

Tal Linzen and Marco Baroni. 2020. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. 2016. Relpron: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018. Neural-davidsonian semantic proto-role labeling. *arXiv preprint arXiv:1804.07976*.