# Structured Neural Networks (II)

## CS 690N, Spring 2018

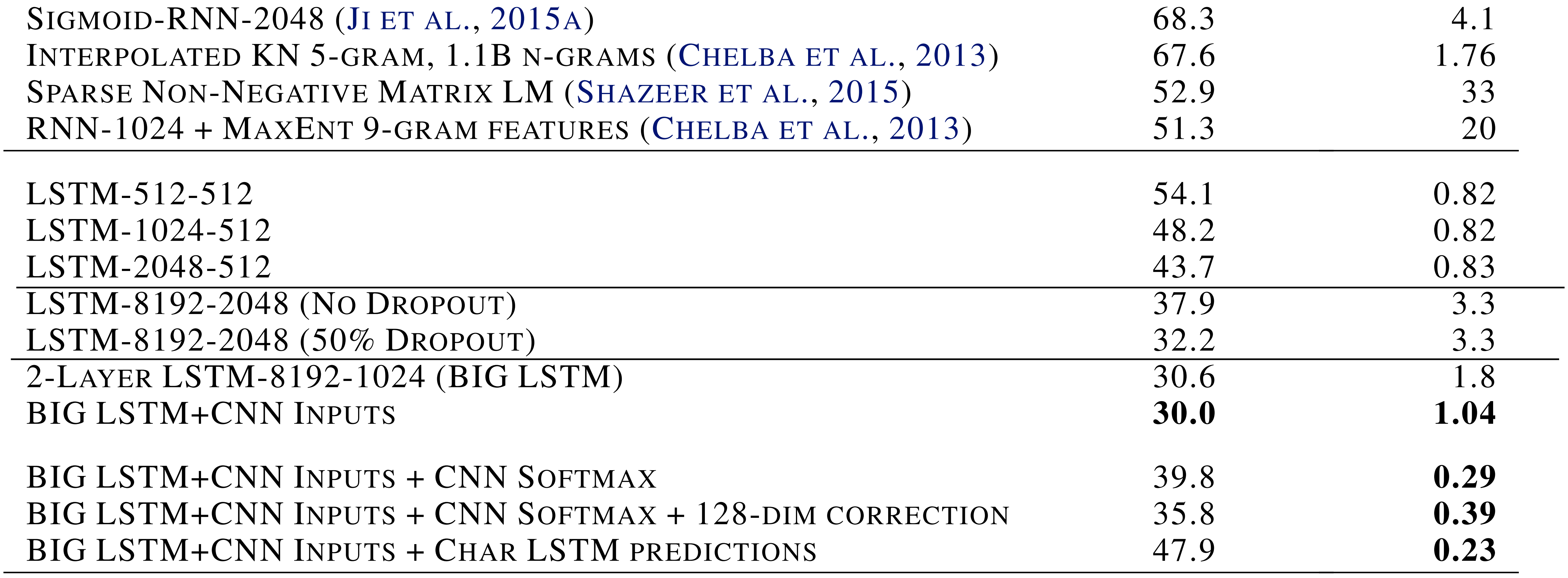
Advanced Natural Language Processing

http://people.cs.umass.edu/~brenocon/anlp2018/

Brendan O’Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

- Some recent large-scale LSTM LM results (V=793471)
*Jozefowicz et al. 2016* https://arxiv.org/pdf/1602.02410.pdf



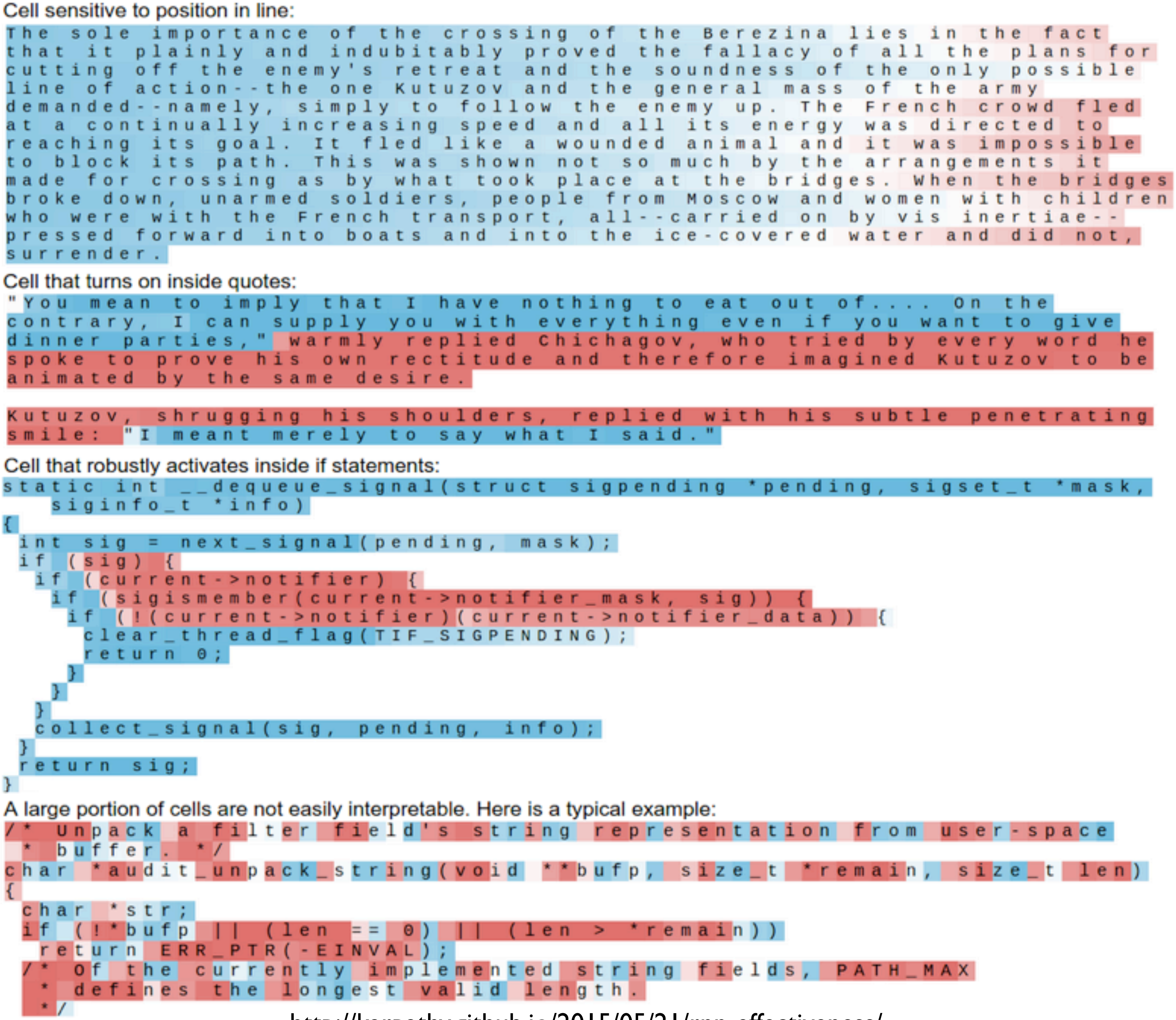
| Model | PPL | Params (billions) |
|---|---|---|
| Sigmoid-RNN-2048 (Ji et al., 2015a) | 68.3 | 4.1 |
| Interpolated KN 5-gram, 1.1B n-grams (Chelba et al., 2013) | 67.6 | 1.76 |
| Sparse Non-Negative Matrix LM (Shazeer et al., 2015) | 52.9 | 33 |
| RNN-1024 + MaxEnt 9-gram features (Chelba et al., 2013) | 51.3 | 20 |
| LSTM-512-512 | 54.1 | 0.82 |
| LSTM-1024-512 | 48.2 | 0.82 |
| LSTM-2048-512 | 43.7 | 0.83 |
| LSTM-8192-2048 (No Dropout) | 37.9 | 3.3 |
| LSTM-8192-2048 (50% Dropout) | 32.2 | 3.3 |
| 2-Layer LSTM-8192-1024 (BIG LSTM) | 30.6 | 1.8 |
| BIG LSTM+CNN Inputs | **30.0** | **1.04** |
| BIG LSTM+CNN Inputs + CNN Softmax | 39.8 | **0.29** |
| BIG LSTM+CNN Inputs + CNN Softmax + 128-dim correction | 35.8 | **0.39** |
| BIG LSTM+CNN Inputs + Char LSTM predictions | 47.9 | **0.23** |

# Softmax alternatives

- Vocabulary softmax is often a bottleneck.

- Hierarchical softmax
- Negative (contrastive) sampling for training
- Character models (?)

# Structure awareness



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
   siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
  if (current->notifier) {
   if (sigismember(current->notifier_mask, sig)) {
    if (!(current->notifier)(current->notifier_data)) {
     clear_thread_flag(TIF_SIGPENDING);
     return 0;
    }
   }
  }
  collect_signal(sig, pending, info);
 }
 return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
 char *str;
 if (!*bufp || (len == 0) || (len > *remain))
  return ERR_PTR(-EINVAL);
 /* Of the currently implemented string fields, PATH_MAX
  * defines the longest valid length.
  */
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

- LSTMs used as a generic, sequence-aware model within language modeling, translation generation, classification and tagging

- Various LSTM-analyzing-text visualizations
  - http://karpathy.github.io/2015/05/21/rnn-effectiveness/
  - http://lstm.seas.harvard.edu/

- Question: can they learn interactions we *know* are in natural language?
  - Thursday: Linzen et al.!

# Syntax in LSTMs

- Can LSTMs capture *natural language* structure?
- Test in different settings (Linzen et al. 2016)
  - Direct supervision (grammatical number prediction)
  - No supervision (LM)

- Subject-Verb agreement on grammatical number

(1) a. The **key is** on the table.
b. *The **key are** on the table.
c. *The **keys is** on the table.
d. The **keys are** on the table.

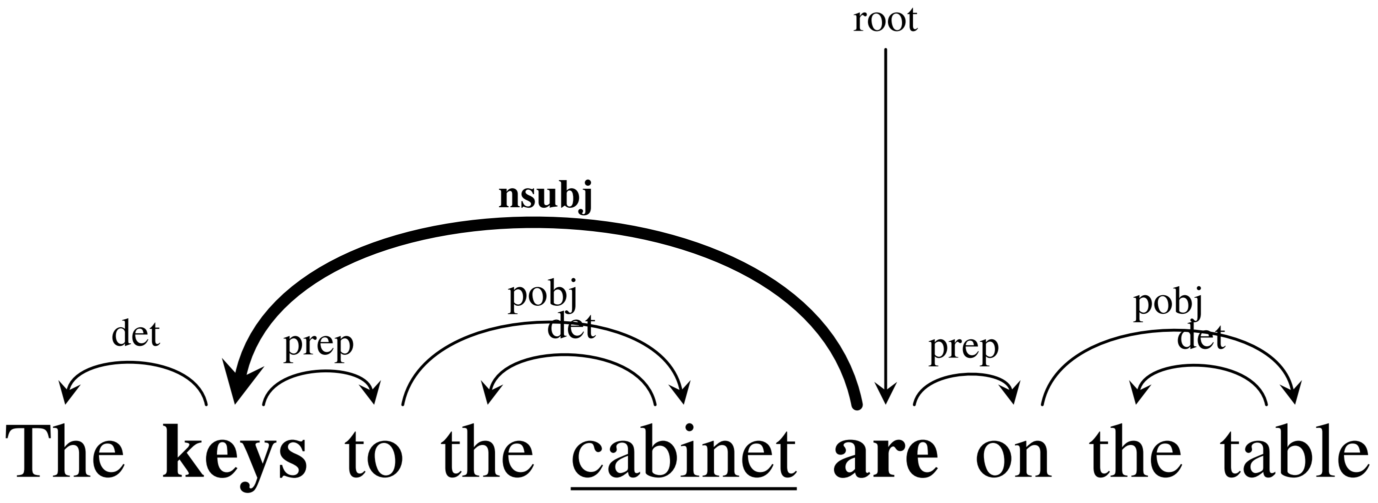
- N-grams can’t capture long-distance dependencies

(2) The **keys** to the cabinet **are** on the table.

(3) The **building** on the far right that’s quite old
and run down **is** the Kilgore Bank Building.

- Use syntactic parser to preprocess data, to generate prediction task setup. (Assumes parser is accurate enough)

> Given a syntactic parse of the sentence and a verb, it is straightforward to identify the head of the subject that corresponds to that verb, and use that information to determine the number of the verb (Figure 1).

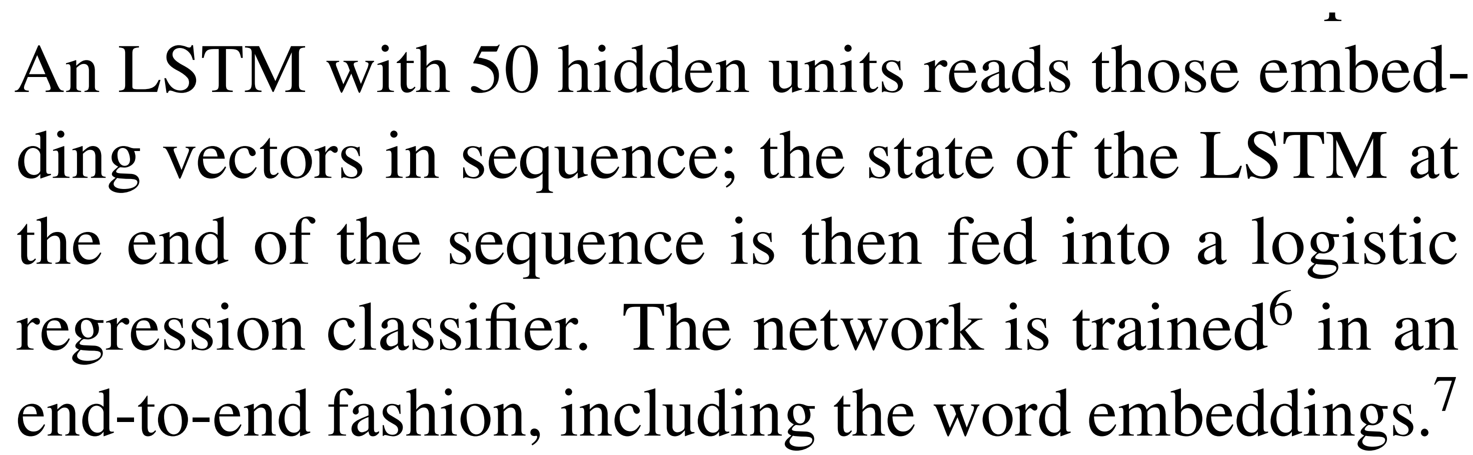


# Number prediction

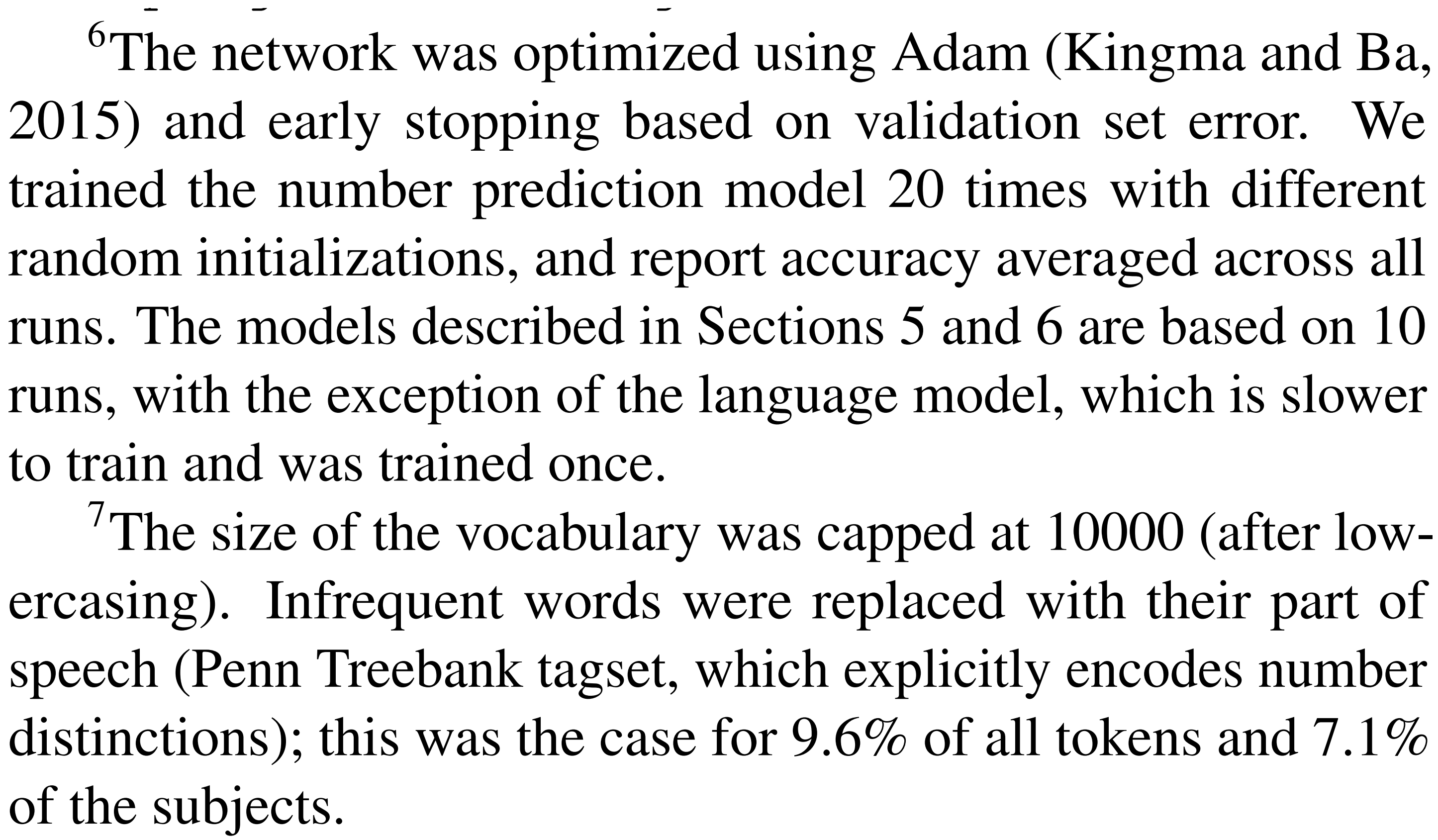
(8) The keys to the cabinet ____

- Task:
  - Predict PLURAL or SINGULAR
  - Needs to learn “subjecthood” and number
  - Unlimited synthetic data (1.3M from Wikipedia: present-tense verb uses)
  - A simple phenomenon that sometimes needs to deal with a little bit of structure
- Models
  - LSTM with 50-dim word embeddings, 50-dim hidden states, last state for classification
  - Noun-only baselines
- Analysis: what affects performance?

# Good reporting of details



An LSTM with 50 hidden units reads those embedding vectors in sequence; the state of the LSTM at the end of the sequence is then fed into a logistic regression classifier. The network is trained[6] in an end-to-end fashion, including the word embeddings.[7]



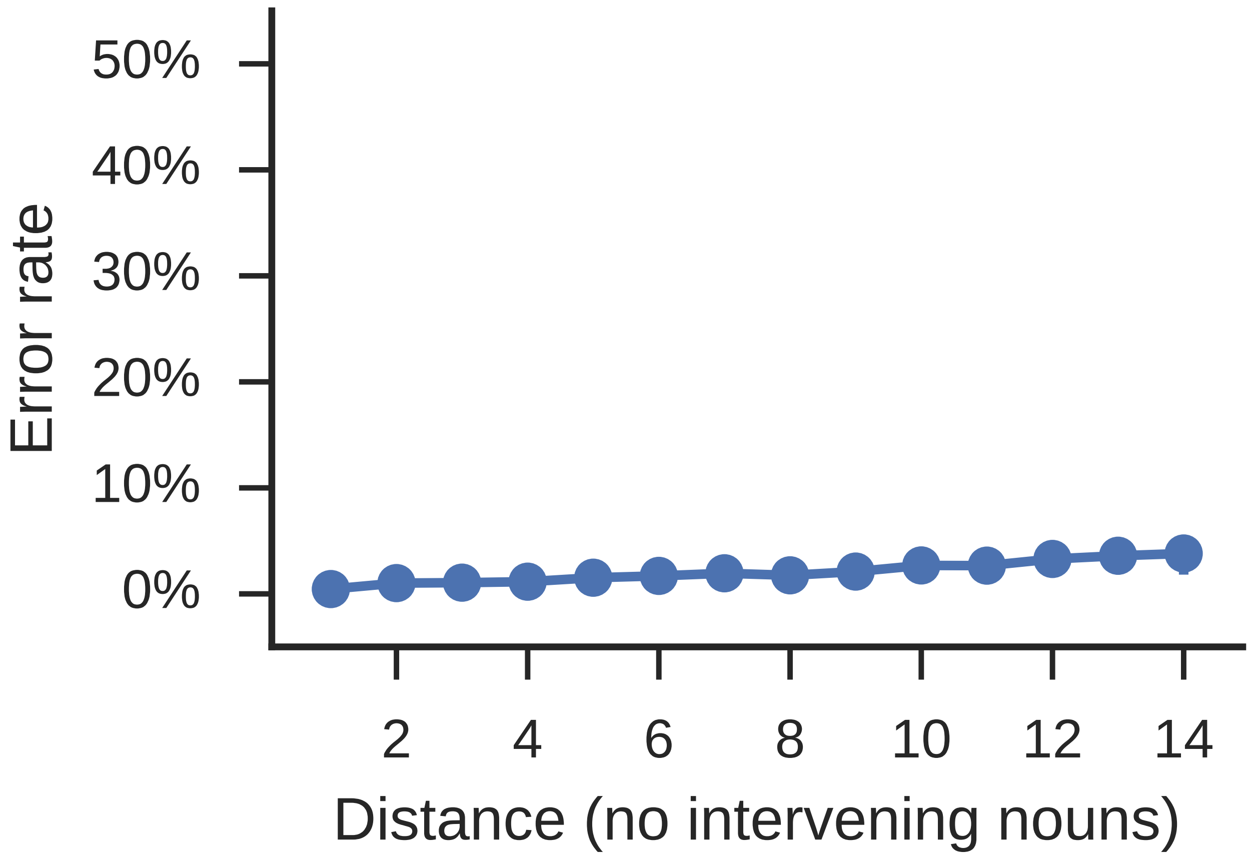
[6] The network was optimized using Adam (Kingma and Ba, 2015) and early stopping based on validation set error. We trained the number prediction model 20 times with different random initializations, and report accuracy averaged across all runs. The models described in Sections 5 and 6 are based on 10 runs, with the exception of the language model, which is slower to train and was trained once.

[7] The size of the vocabulary was capped at 10000 (after lowercasing). Infrequent words were replaced with their part of speech (Penn Treebank tagset, which explicitly encodes number distinctions); this was the case for 9.6% of all tokens and 7.1% of the subjects.

# What affects performance?

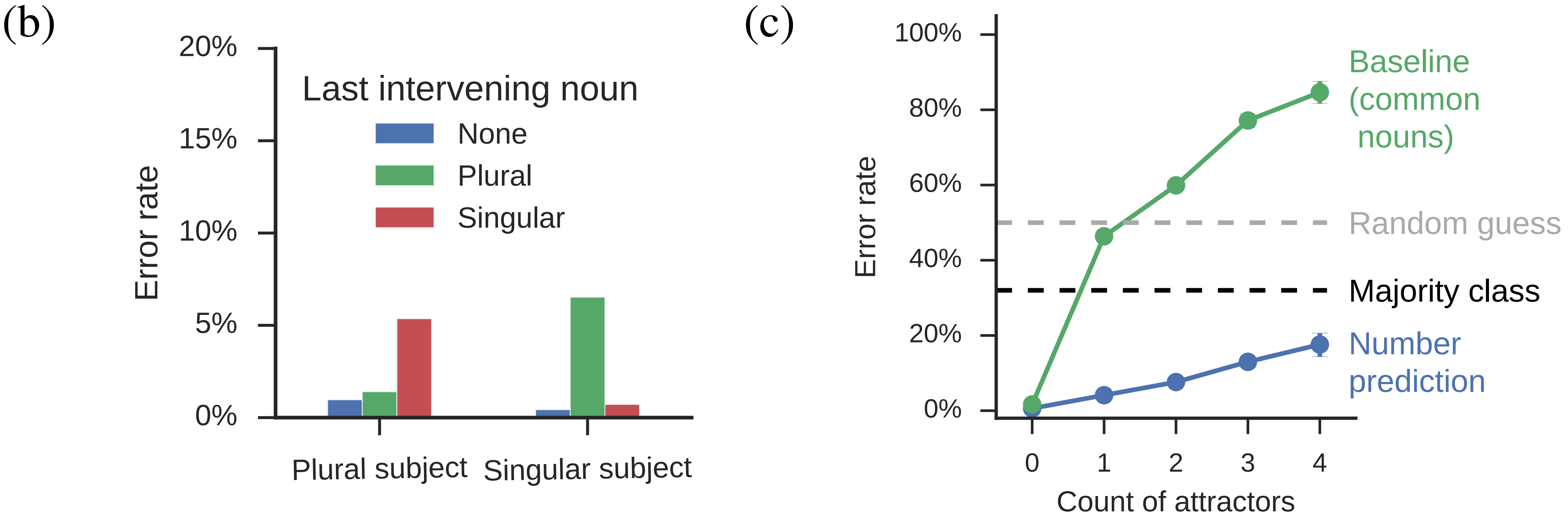
- Distance?

(a)



# What affects performance?

- Agreement attractors: do intervening nouns distract the model?



- Yes, but not fatal -- especially compared to guessing and if deprived of function words
- Multiple intervening nouns: "homogeneous intervention" of same number
  - Yes: The **roses** in the vase by the door **are** red.
  - No: The **roses** in the vase by the chairs **are** red.

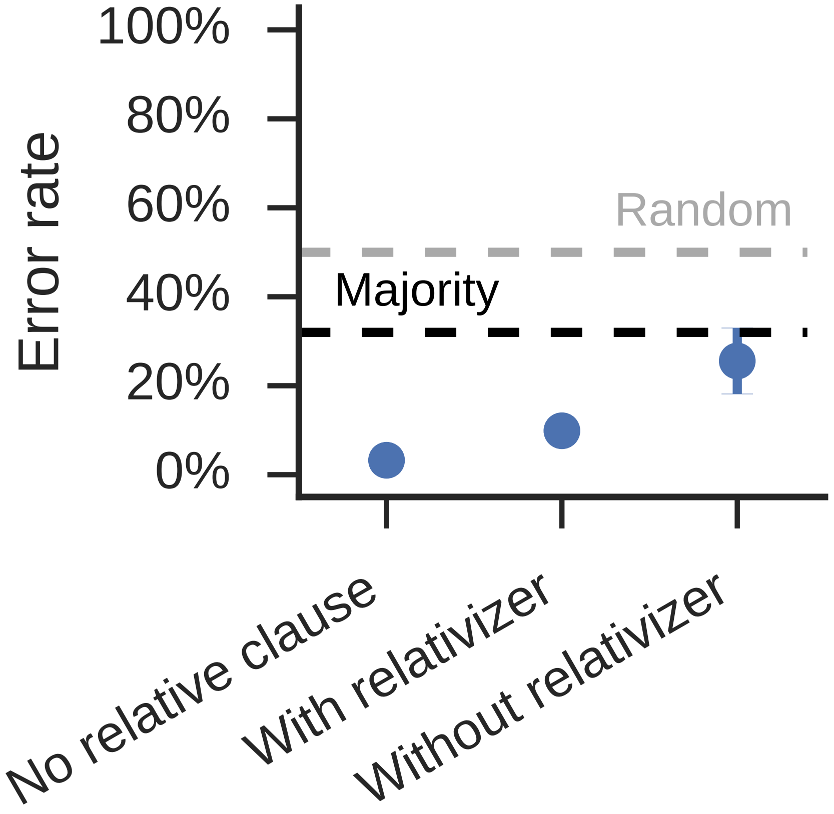
# What affects performance?

- Intervening nouns when in relative clauses? Challenging:
  - The RC has its own subject-verb pair with their own grammatical number
  - It may or may not have an explicit *relativizer* word

(11) The **landmarks** this article lists here **are** also run-of-the-mill and not notable.

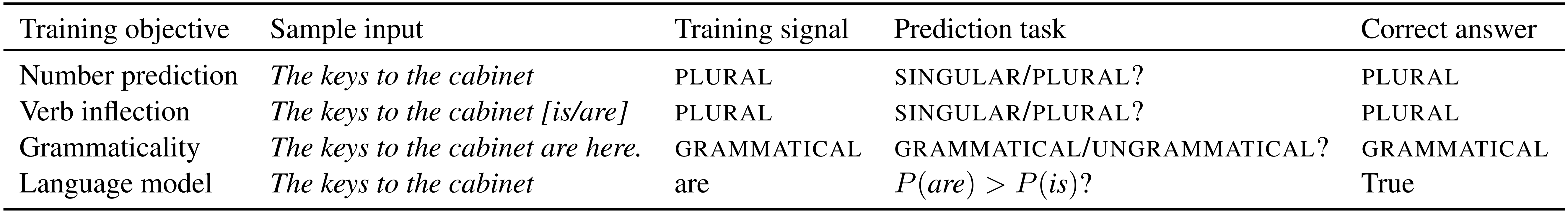
(12) The **landmarks** *that* this article lists here **are** also run-of-the-mill and not notable.

(d)



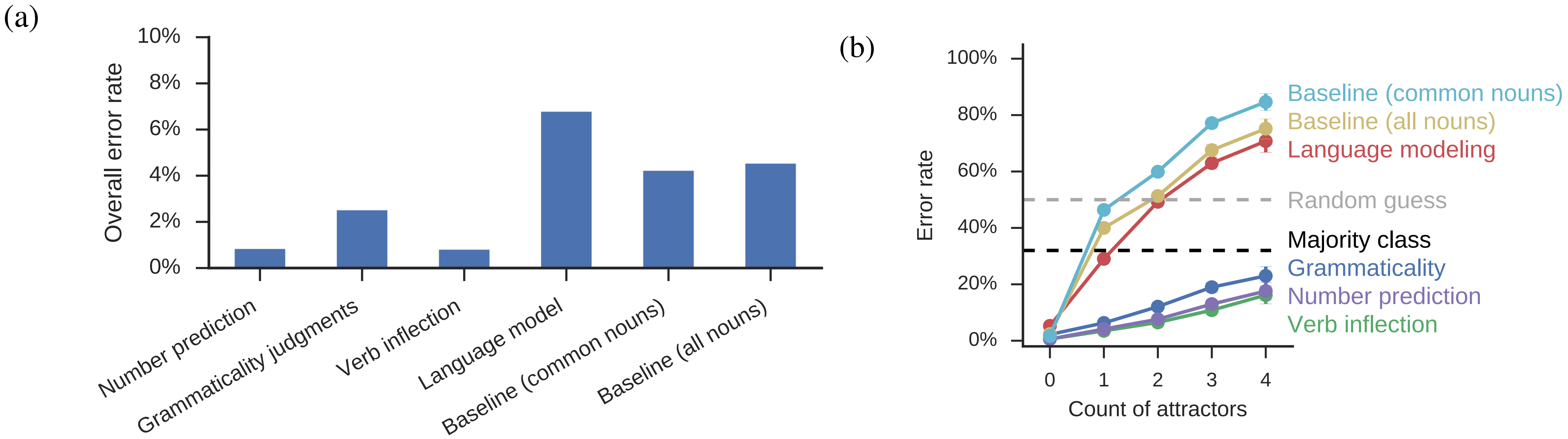
# Language model

- Does an LSTM LM implicitly learn these syntactic rules?
  - Assess number prediction by comparing e.g. P(writes | ...) vs. P(write | ...)



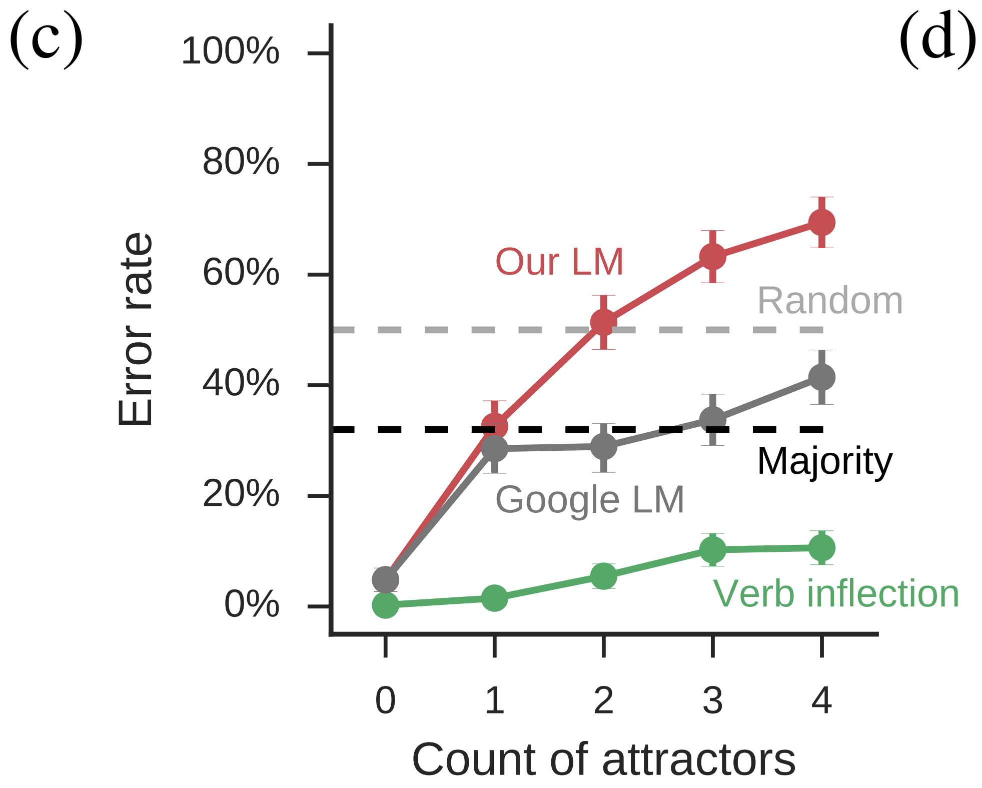
| Training objective | Sample input | Training signal | Prediction task | Correct answer |
|---|---|---|---|---|
| Number prediction | *The keys to the cabinet* | PLURAL | SINGULAR/PLURAL? | PLURAL |
| Verb inflection | *The keys to the cabinet [is/are]* | PLURAL | SINGULAR/PLURAL? | PLURAL |
| Grammaticality | *The keys to the cabinet are here.* | GRAMMATICAL | GRAMMATICAL/UNGRAMMATICAL? | GRAMMATICAL |
| Language model | *The keys to the cabinet* | are | $P(are) > P(is)$? | True |

Table 1: Examples of the four training objectives and corresponding prediction tasks.



# Language model

- Even large-scale LM ("Google LM", trained on 1B words) still lags the more directly supervised model



# Hidden unit visualizations

- cherry picking?

# Quetions: Scope

(3) Compared to our other readings which examined the perplexity of a corpus, this paper's model seem severely constrained to binary classification of subject-verb plurality. It's cool that we can do this with a LSTM, but the scope of the paper is not that ambitious. Creating an LSTM network to determine if a subject and verb agree or not is markedly less impressive than using an LSTM network to determine if the entire sentence is grammatically correct.

The weakness of the paper is that it doesn't provide any significant technical contribution to existing work. It is just a series of experiments. There is no concrete theory or contribution.

(5) What would happen if we tried something similar to our first homework assignment? That is, what if we took the sentences from the corpus, introduced a corresponding sentence for each with a jumbled word order, and trained the LSTM to identify the grammatically correct sentences? Could this function as an alternative to the ngram method?

One was the way they split their training, validation and test sets. It seemed very unusual and it seemed they mostly did it to speed up training and experiments.

Second, I do not think they gave a very good reason for why they choose subject-verb number agreement out of all the other possible structure sensitive dependencies found in text.

# Human performance?

Human-error scores for this would be extremely helpful in knowing if the proposed tasks and objectives can give some extra information to the LM that even humans fail to capture.

One direction (which I think may have been taken; I think someone presented something like this at SCIL) would be to compare the neural network's predictions with human performance on agreement attraction cases. As the authors note, humans make a lot of mistakes in similar contexts, and there is a lot of experimental work on the topic. Comparing the performance of humans and the neural network and seeing whether they make mistakes in similar contexts could help us understand whether the neural network is learning in a human-like way.

1. The results are interesting, but where does this take us? I'm confused about what the authors think LSTMs can achieve practically. In the beginning of the paper, they mention that RNNs are used in parsing, translation and other tasks, which is why this study is relevant, but does the conclusion advocate for or against their use?

2. The Subject-Verb-Object structure is only found in English and similar languages right? What about a language like Irish where the verb would precede the subject? How would that affect their experiments? For instance if there are multiple nouns following a verb, would it be possible to identify the subject based on the number of the verb? I'm not able to come up with an example though.

[The authors] stated that syntactic parsing driven approaches are prone to failure. However, we may think of adopting syntactic parser as auxiliary feature for the agreement prediction model of LSTM. Recently, some [research has] reported that some NLP tasks can be boosted by using syntactic structure information.

# More/explicit hierarchy?

I would be really interested to see how do architectures which explicitly account for the compositional nature of language by having a hierarchical structure, compare on tasks like the verb number prediction task. I am really curious to see if architectures like those proposed in Tai et al's 'Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks' are able to work as well without explicit supervision and really harness the grammatical structure of the language.

- Excellent illustration of model analysis
    - Analyze model performance with respect to research questions
    - Break down errors by properties of examples
    - Visualizations
    - Scientific understanding of computational linguistics

# Questions

- LSTMs can impressively learn longer-range interactions in real natural language data
  - Previous work: artificial languages
- Total unsupervised learning not as good as supervised syntactic signal: why??

- Is there a model class such at simple LM training will capture all of language?
- What supervision do we need for good NLP systems?