Structured Neural Networks (II)

CS 690N, Spring 2017

Advanced Natural Language Processing

http://people.cs.umass.edu/~brenocon/anlp2017/

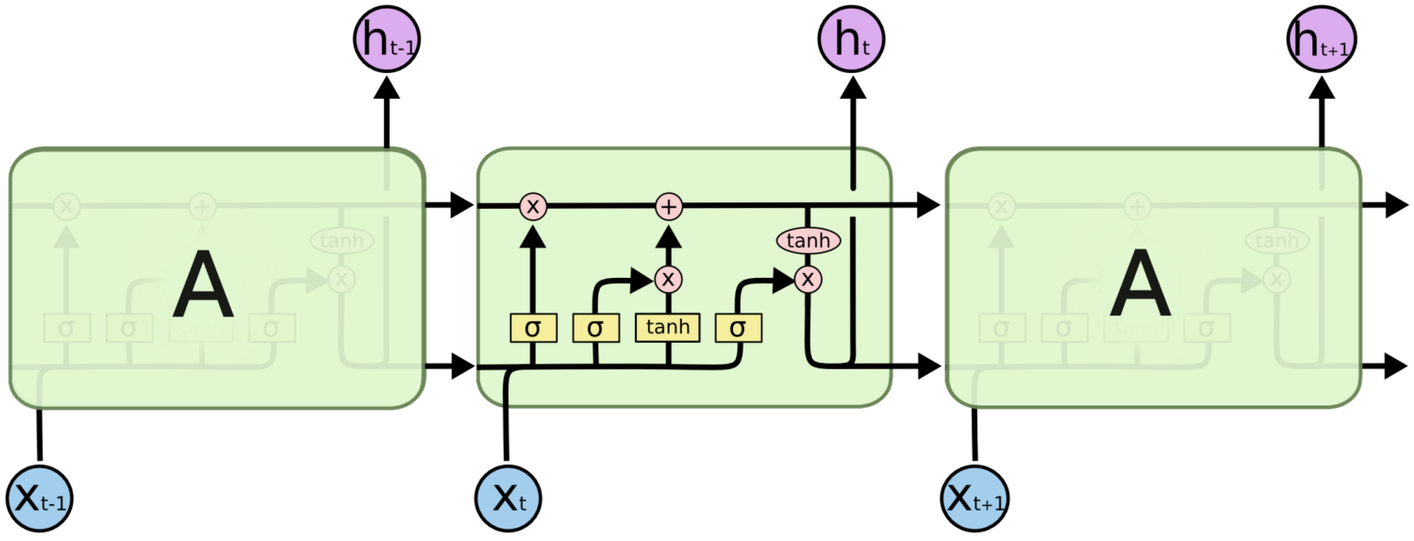
Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

LSTM (Long short-term memory)

- Goal: be able to "remember" for longer distances
- Augment individual timesteps with a number of specialized vectors and gating functions
  - c: Memory component (a.k.a. *cell*)
  - h: Hidden state
  - f,i,o: Forget, Input, Output
  - g: proposed new state. f,i,o decide how much to accept it.
- (See GRU for a simpler, more intuitive model that does the same thing. But LSTM seems to be the most common RNN currently.)



$$\mathbf{s_j} = R_{\text{LSTM}}(\mathbf{s_{j-1}}, \mathbf{x_j}) = [\mathbf{c_j}; \mathbf{h_j}]$$

$$\mathbf{c_j} = \mathbf{c_{j-1}} \odot \mathbf{f} + \mathbf{g} \odot \mathbf{i}$$

$$\mathbf{h_j} = \tanh(\mathbf{c_j}) \odot \mathbf{o}$$

$$\mathbf{i} = \sigma(\mathbf{x_j}\mathbf{W^{xi}} + \mathbf{h_{j-1}}\mathbf{W^{hi}})$$

$$\mathbf{f} = \sigma(\mathbf{x_j}\mathbf{W^{xf}} + \mathbf{h_{j-1}}\mathbf{W^{hf}})$$

$$\mathbf{o} = \sigma(\mathbf{x_j}\mathbf{W^{xo}} + \mathbf{h_{j-1}}\mathbf{W^{ho}})$$

$$\mathbf{g} = \tanh(\mathbf{x_j}\mathbf{W^{xg}} + \mathbf{h_{j-1}}\mathbf{W^{hg}})$$

$$\mathbf{y_j} = O_{\text{LSTM}}(\mathbf{s_j}) = \mathbf{h_j}$$

Christopher Olah: Understanding LSTM Networks
`colah.github.io/posts/2015-08-Understanding-LSTMs/`

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/

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.

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)
- 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

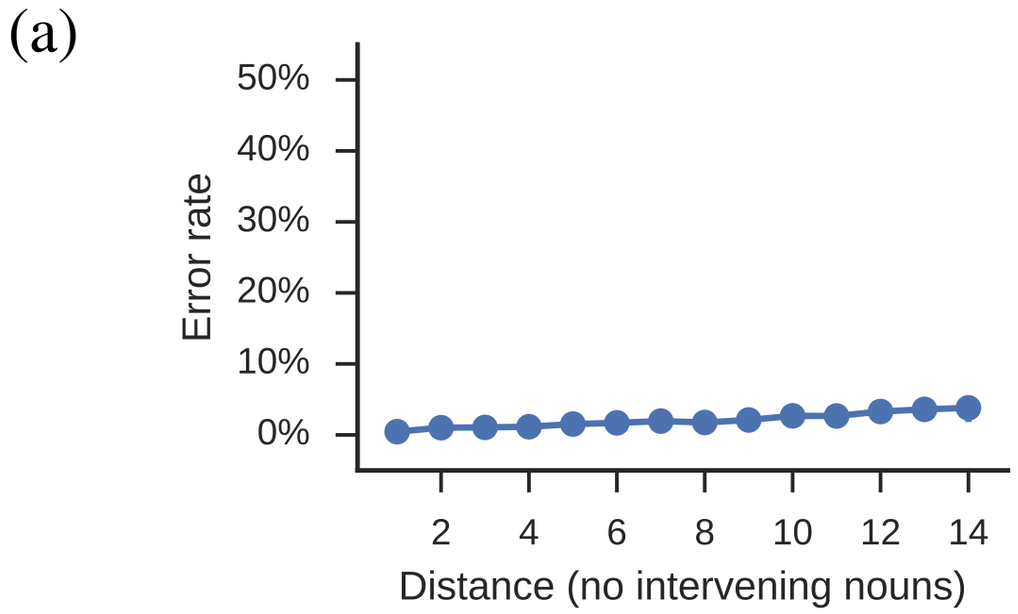
An LSTM with 50 hidden units reads those embedding vectors in sequence; the state of the LSTM at the end of the sequence is then fed into a logistic regression classifier. The network is trained[6] in an end-to-end fashion, including the word embeddings.[7]

[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?

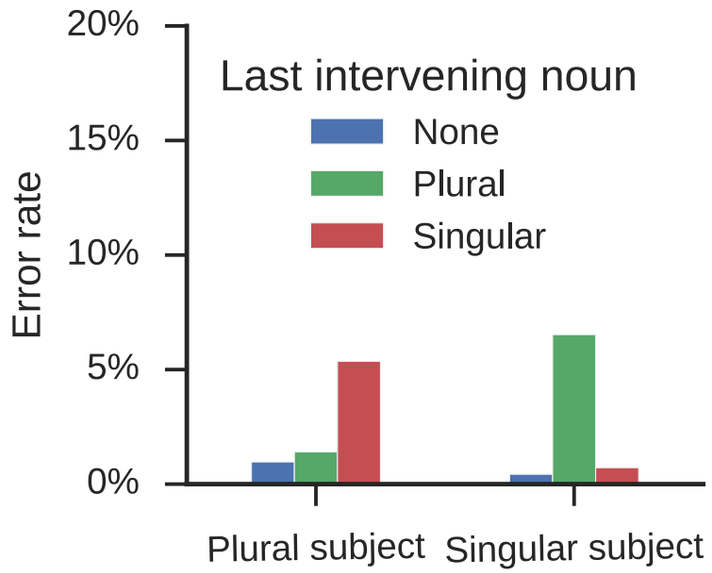
- Distance?



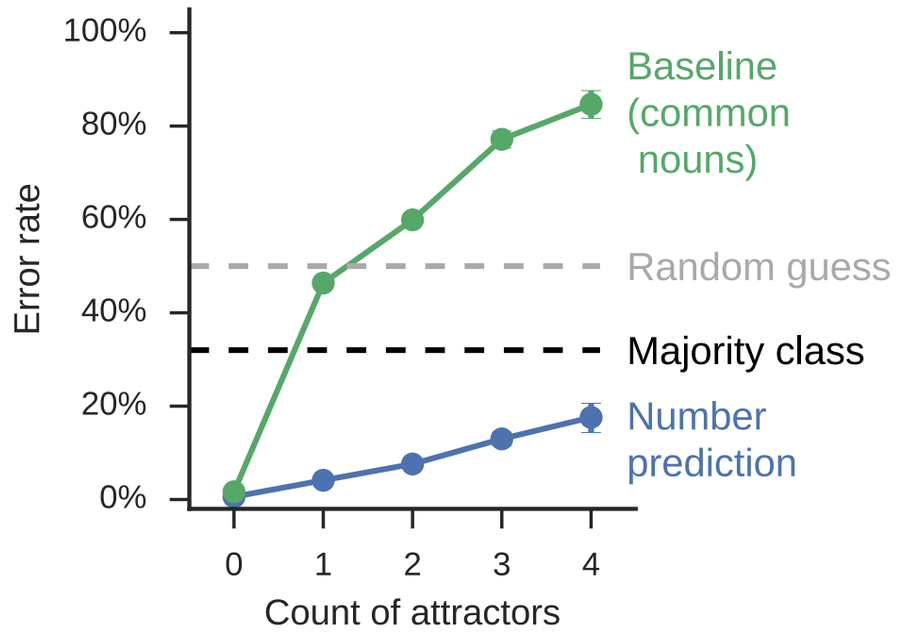
What affects performance?

- Agreement attractors: do intervening nouns distract the model?

(b)



(c)



- 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.

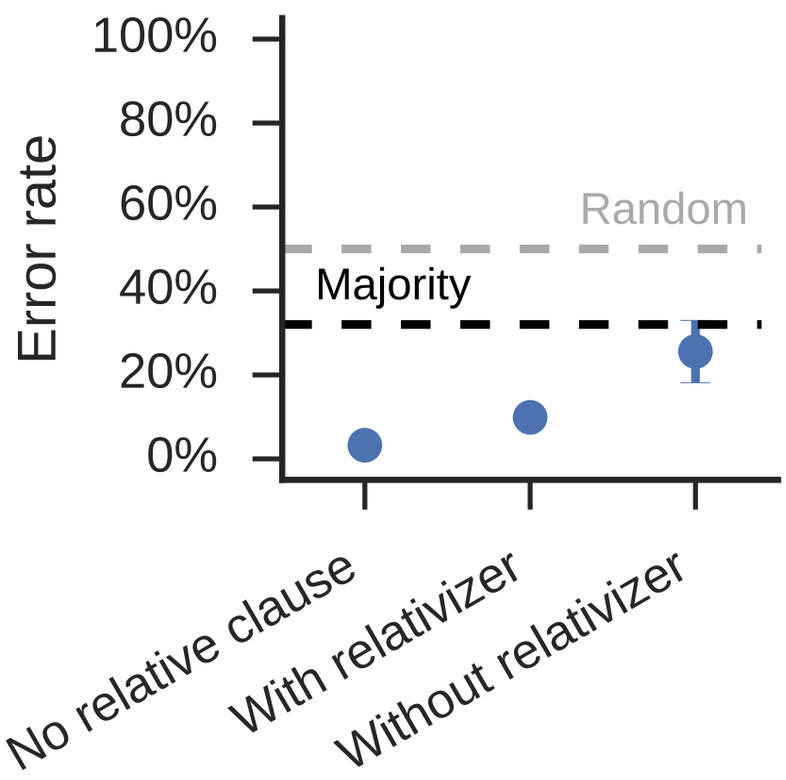
What affects performance?

- 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.

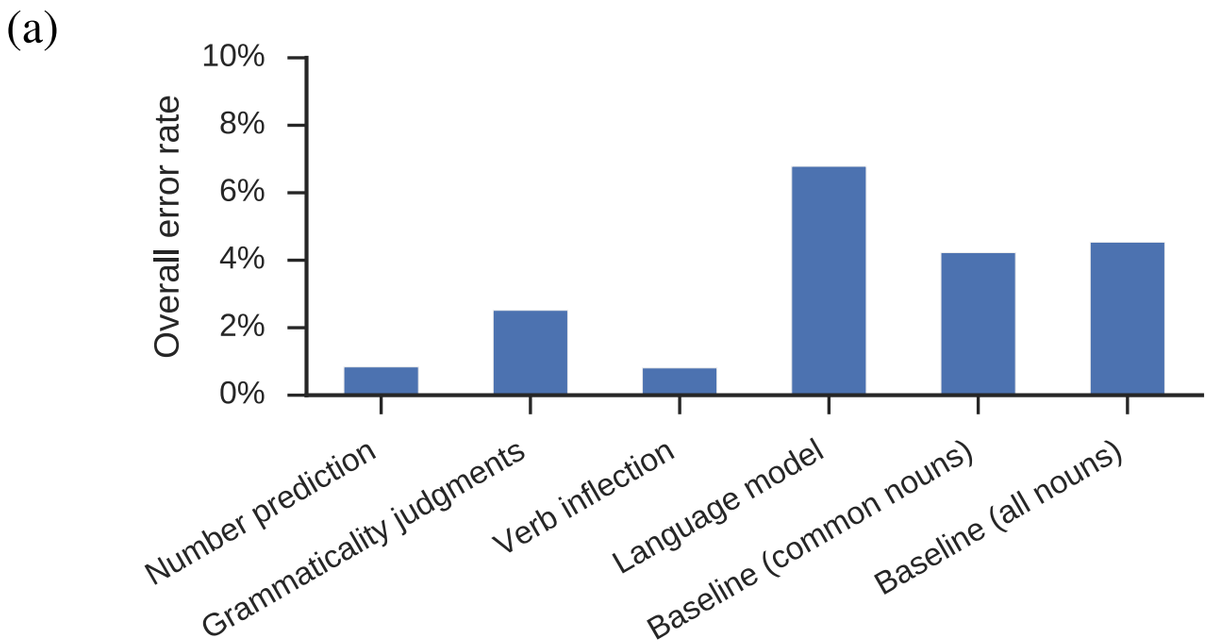
(12) The **landmarks** *that* this article lists here **are** also run-of-the-mill and not notable.

(d)



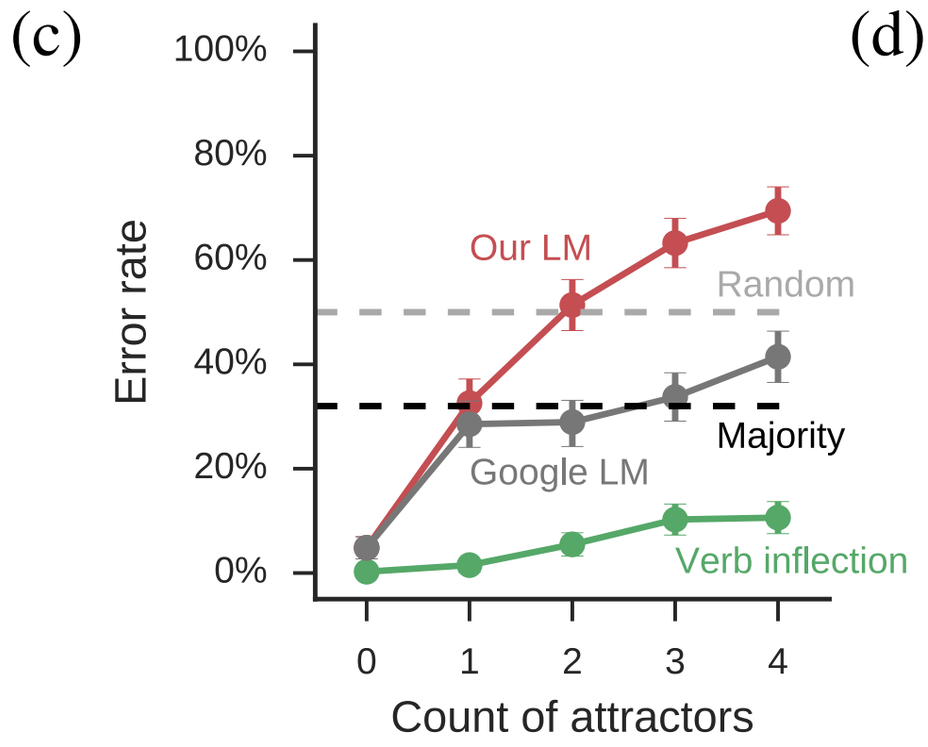
Language model

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



Language model

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



Conclusions

- 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

- 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

Recursive Neural Networks

- (Whiteboard)

- Reference for all-nodes supervision:
  Stanford Sentiment Treebank
  https://nlp.stanford.edu/sentiment/treebank.html