RL LAB

Capturing Pragmatic Knowledge in Article Usage Prediction using LSTMs

Jad Kabbara, Yulan Feng & Jackie C.K. Cheung

McGill University

December 16, 2016

RNNs and NLP

- RNNs have been successful in many NLP tasks recently:
  - Language modeling (Mikolov et al., 2010)
  - Machine translation (Sutskever et al., 2014; Cho et al. 2014, Bahdanau at al. 2015)
  - Dependency parsing (Dyer et al., 2015)

Why are RNNs powerful?

- Representational power
- Long range dependencies
- The possibility of using pre-trained word embeddings (e.g., word2vec)

Applications in Computational Discourse and Pragmatics

- Discourse parsing (Kalchbrenner and Blunsom, 2013; Ji and Eisenstein, 2014; Li et al., 2014)
- Implicit discourse relation detection (Ji and Eisenstein, 2015)
- Building distributed representations of linguistic units larger than a sentence (Le and Mikolov, 2014; Li et al., 2015)
- Predicting function and content words (Hill et al., 2016)

Our Work

- Our interest in this work is to examine whether these purported benefits of RNNs can be used to improve the modelling of pragmatic effects in language.
- Task: Definiteness prediction

Definiteness Prediction

- Definition: The task of determining whether a noun phrase should be definite or indefinite.
- One case (in English): Predict whether to use a definite article (the), indefinite article (a(n)), or no article at all.
- Applications: MT, summarization, L2 grammatical error detection and correction.

Why is it interesting linguistically?

- Both contextual and local cues are crucial to determining the acceptability of a particular choice of article.

Contextual cues

- “The” asserts existence and uniqueness of entity in context (Russell, 1905)
- Anaphoric nature; ability to trigger a presupposition about the existence of the NP in the discourse context (Strawson, 1950)

Contextual Cues

- Role of factors such as discourse context, familiarity, and information status
- Example:

  A/#the man entered the room. The/#a man turned on the TV.

Non-Context-Dependent Factors

- May block articles:

  Demonstratives (e.g., this, that), Certain quantifiers (e.g., no), Mass nouns (e.g., money)
- Conventions for named entities (which article to use, or whether to use an article at all):
  - The United Kingdom (definite article required)
  - Great Britain (no article allowed).

Our Questions

- How much linguistic “knowledge” do we need to explicitly encode in a system that predicts definiteness?
- Can a statistical learner, such as RNNs, learn interpretable complex features for this prediction task?
- Can RNNs pick up on local and non-local cues?

Previous Work

- Rely heavily on hand-crafted linguistic features:
  - Knight and Chander, 1994; Minnen et al., 2000; Han et al., 2006; Gamon et al., 2008
  - Turner and Charniak (2007) trained a parser-based language model on the WSJ and North American News Corpus.

Previous State-of-the-Art

- De Felice (2008): Learn a logistic regression classifier using 10 types of linguistic features
  - Example: Pick ***(the?)*** juiciest apple on the tree.

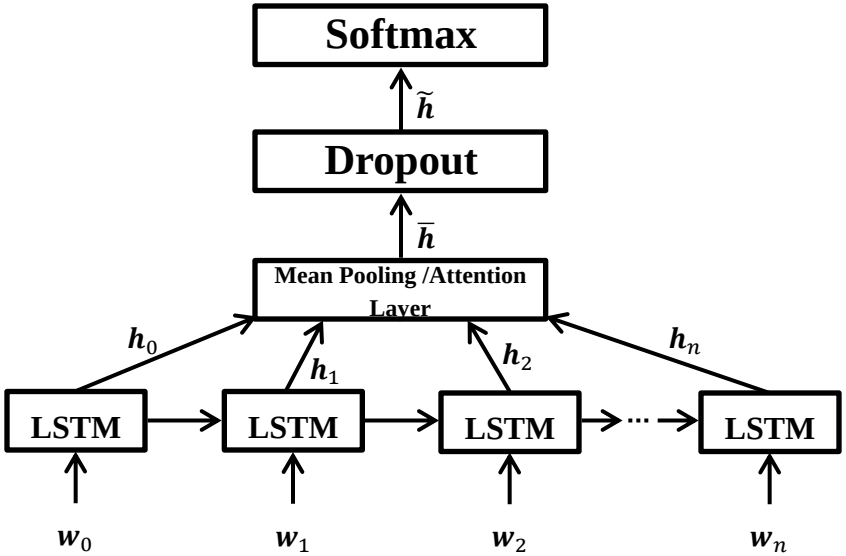
Head noun	**'apple'**
Number	singular
Noun type	count
Named entity?	no
WordNet category	food, plant
Prep modification?	yes, 'on'
Object of Prep?	no
Adjective modification?	yes, 'juicy'
Adjective grade	superlative
POS ±3	VV, DT, JJS, IN, DT, NN

Our Approach: Deep Learning

- Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber 1997)-based RNN
- Two variants:
  - Vanilla model, Attention-based model
- Different initializations of word embeddings:
  - Random initialization, Word2vec, GloVe vectors

Our Approach: Deep Learning

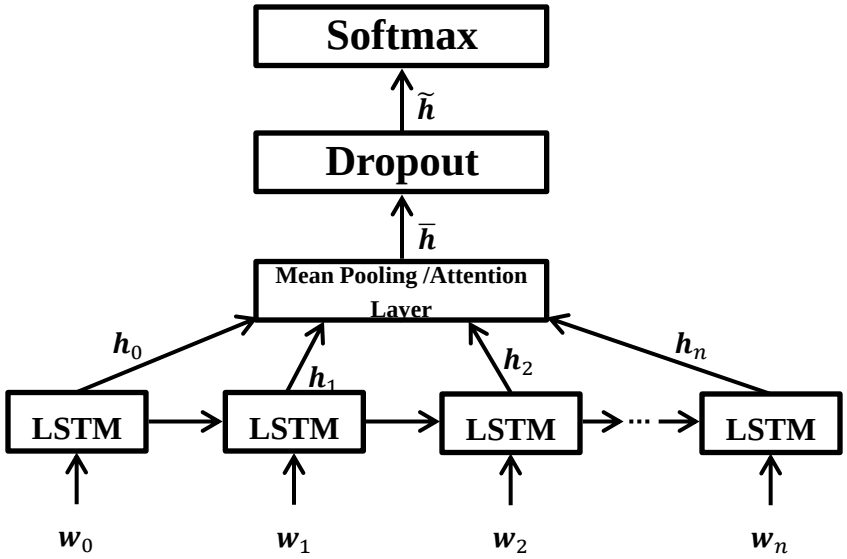
- LSTM-based recurrent neural network



Brief Detour: Attention

- Loosely inspired by theories of human visual attention in which specific regions of interest have high focus compared to other regions.
- Adopted in Neural Networks research (e.g., Bahdanau et al., 2014)
- Interpretability

Brief Detour: Attention



- Vanilla model:
  - $\bar{h} = \frac{1}{n}\sum_i h_i$
- Attention-based model:
  - $c_i = \tanh(Wh_i + b)$
  - $a_i = \frac{e^{c_i}}{\sum_j e^{c_j}}$
  - $\bar{h} = \sum_i a_i h_i$

Local Context

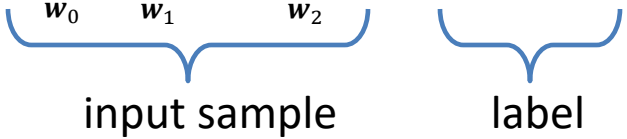
- Sample configuration for local context:
  - The set of tokens from the previous head noun of a noun phrase up to and including the head noun of the current noun phrase.

Local Context

- Example: For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.

Local Context

- Example: For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.
  - For six years – '**none**'



Local Context

- Example: For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.
  - For six years – '**none**'
  - T. Marshall Hahn Jr – '**none**'

input sample label

Local Context

- Example: For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.
  - For six years – '**none**'
  - T. Marshall Hahn Jr – '**none**'
  - has made corporate acquisitions – '**none**'

input sample | label

Local Context

- Example: For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.
  - For six years – '**none**'
  - T. Marshall Hahn Jr – '**none**'
  - has made corporate acquisitions – '**none**'
  - in ~~**the**~~ George Bush mode – '**the**'.

Extended Context

- In addition to the “local context” tokens, add the preceding tokens such that the total number of tokens is a fixed number *N*.

Extended Context

- Consider 3rd sample of previous example(N=50):
  For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.

… For six years, T . Marshall Hahn Jr. has made corporate acquisitions - **none**

w_0 w_1 … … w_{46} w_{47} w_{48} w_{49}

"extended context" input sample

label

Extended Context

- Consider 3rd sample of previous example(N=50):
  For six years, T. Marshall Hahn Jr. has made corporate acquisitions in the George Bush mode: kind and gentle.

… For six years, T . Marshall Hahn Jr. has made corporate acquisitions - **none**

w_0 w_1 … … w_{46} w_{47} w_{48} w_{49}

"extended context" input sample

label

Word Embeddings

- Random Initialization
- Word2vec (Mikolov et al., 2013)
  - 300 dimensions
  - Trained on the Google News corpus(~100 billon words)
- GloVe vectors (Pennington et al., 2014)
  - 300 dimensions
  - Trained on the Common Crawl corpus (~840 billon words)

Experiment setup

- Datasets:
  - Penn Treebank (PTB) – WSJ articles
    - ∼223k samples for training
    - ∼18k samples for development
    - ∼22k samples for testing

Model Comparison

- Baseline: Label all noun phrases with the most frequent class of none
- LogReg (de Felice, 2008)
- LSTM model (Attention? Context? Word embeddings? POS tags?)

Results: Classification Accuracy

Method	Accuracy (%)
None-class baseline	67.70
LogReg	93.07
Best performing LSTM	**96.63**

Results: Classification Accuracy

Method	Accuracy (%)			
None-class baseline	67.70			
LogReg	93.07			
	Initialization	POS	Local context	Extended context
LSTM	Random		83.94	95.82
LSTM	Word2vec		84.91	96.40
LSTM	GloVe		85.35	96.37

Results: Classification Accuracy

Method	Accuracy (%)			
None-class baseline	67.70			
LogReg	93.07			
	Initialization	POS	Local context	Extended context
LSTM	Random	– POS	83.94	95.82
LSTM	Word2vec	– POS	84.91	96.40
LSTM	GloVe	– POS	85.35	96.37
LSTM	Random	+ POS	94.11	95.95
LSTM	Word2vec	+ POS	94.50	96.20
LSTM	GloVe	+ POS	94.64	96.38

Results: Classification Accuracy

Method	Accuracy (%)			
None-class baseline	67.70			
LogReg	93.07			
	Initialization	POS	Local context	Extended context
LSTM	Random	– POS	83.94 - 83.96	95.82 - 96.08
LSTM	Word2vec	– POS	84.91 - 84.93	96.40 - 96.53
LSTM	GloVe	– POS	85.35 - 85.75	96.37 - 96.43
LSTM	Random	+ POS	94.11 - 94.12	95.95 - 96.08
LSTM	Word2vec	+ POS	94.50 - 94.52	96.20 - 96.25
LSTM	GloVe	+ POS	94.64 - 94.67	96.38 - **96.63**

Results: Named Entities

Method	Test Set Accuracy (%)	
	Named Entities (N = 5100)	Non-Named Ent. (N = 16579)
None-class Baseline	86.98	61.76
LogReg	97.27	91.77
Local LSTM+a + GloVe + POS	**98.88**	93.44
Extended LSTM+a + GloVe + POS	97.62	**96.48**

Context Analysis

- Compare the best performing LSTM model that uses local context to the best performing LSTM model that uses “extended context”.
- Investigate 200 samples out of the 957 samples that were incorrectly predicted by the former but correctly predicted by the latter.

Context Analysis

- Group samples in two categories:
  - Simple cases where the decision can be made based on the noun phrase itself (e.g., fixed expressions, named entities)
  - Complex cases where contextual knowledge involving pragmatic reasoning is required (bridging reference, entity coreference involving synonymy)

Context Analysis

	Simple Cases		Complex Cases	
	Fixed Expressions	Duplication of the head noun	Synonyms	Needs semantic understanding
	86	6	8	100
Total	92		108	

Attention-based Analysis

Some snippets of text showing samples that were *correctly* predicted by the model using extended context but *incorrectly* predicted by the model using local context

Attention-based Analysis

... net income for the third quarter of 16.8 million or 41 cents a share reflecting [a] broad-based improvement in the company's core businesses. Retail profit surged but the company *[sic]* it was only a modest contributor to third-quarter results. A year ago, net, at **the** New York investment banking firm ...

Note: Underlined words received the highest attention weights

Attention-based Analysis

... companies. In a possible prelude to the resumption of talks between Boeing Co. and striking Machinists union members, a federal mediator said representatives of the two sides will meet with him tomorrow. It could be a long meeting or it could be a short one, said Doug Hammond, the mediator ...

Note: Underlined words received the highest attention weights

Conclusion

- State of the art for article usage prediction
  - LSTM networks can learn complex dependencies between inputs and outputs for this task.
  - Explicitly encoding linguistic knowledge doesn't seem to hurt, but it doesn't help much either.
  - Providing more context improves the performance

Future Work

- Interesting applications in L1 vs L2 English
- Further experiments on predicting other linguistic constructions involving contextual awareness and presupposition.

Thank you !! ☺