Models for Named Entity Recognition

Who Am I?

Data Scientist at Space Jam Data working on NLP problems in short unstructured text.

Background in Applied Math & Stats.

Interested in connecting concepts across domains using written language.

What is this talk about?

- Named Entity Recognition (NER)
- Quickly introduces CRF as a model for NER
- How can Deep Learning help?
- Build BI-LSTM-CRF with keras

What is Named Entity Recognition?

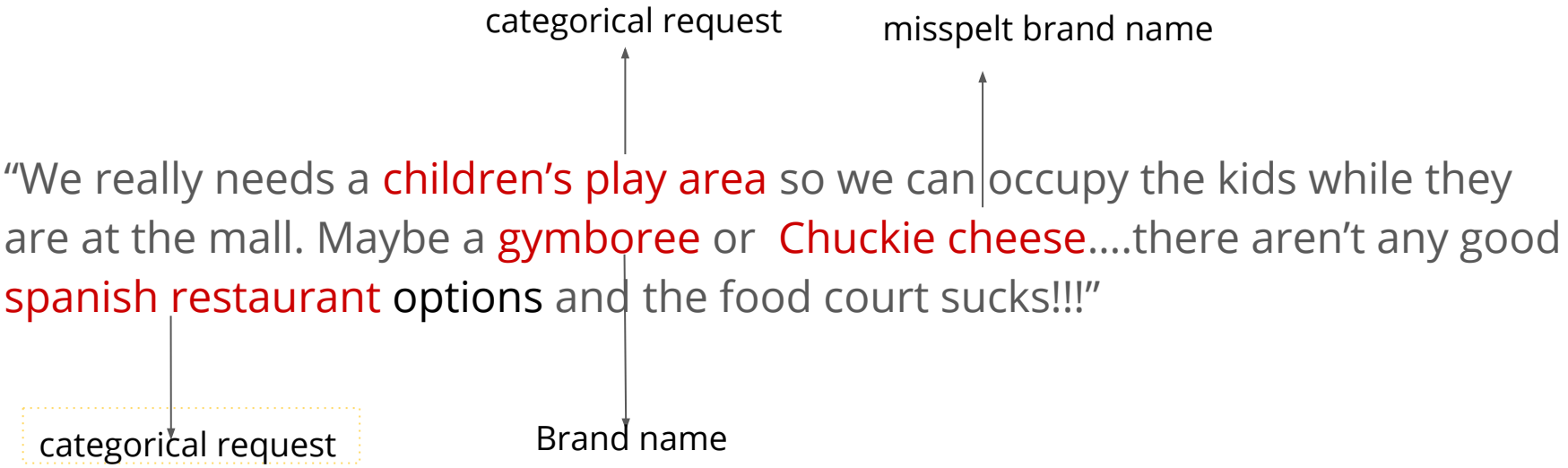
Token Wiki definition

“Named-entity recognition (**NER**) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.” - Wikipedia

Given some text..

“We really needs a children’s play area so we can occupy the kids while they are at the mall. Maybe a gymboree or Chuckie chesse....there aren’t any good restaurant options and the food court sucks!!!”

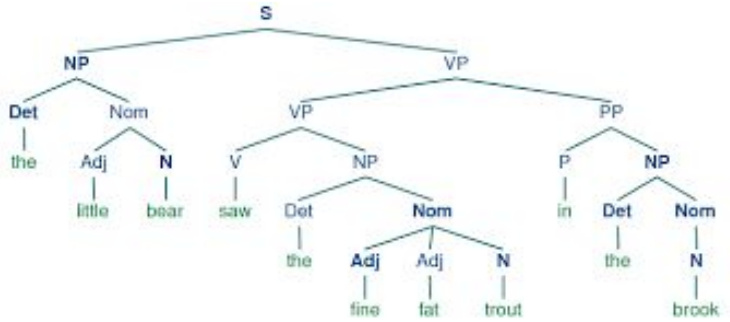
We would like to label



Structured Prediction

Outputs are structured objects such as: sequences, strings, trees.

DET	NN	NN	PREP	ADJ	NN
The	Orwellian	days	of	old	children.



paraphrasing http://mlg.eng.cam.ac.uk/mlss09/mlss_slides/Hoffman_1_2.pdf slide 5.

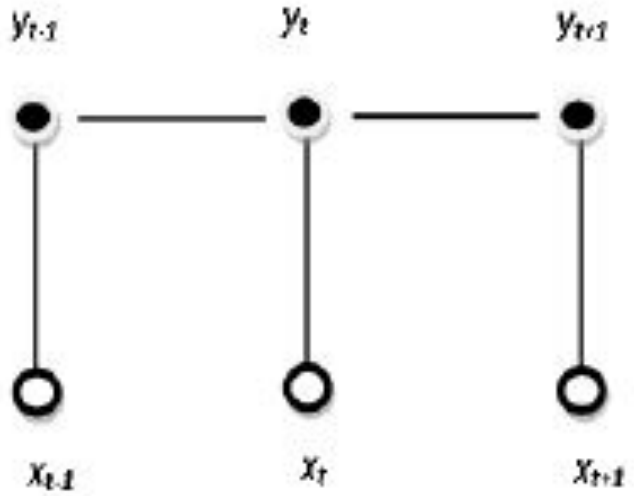
Models

- Conditional Random Fields (CRFs)
- Structured SVMs
- Max-Margin Markov Model (M^3 Networks)
- HIdden Markov Models (HMMs)
- Others

What is a Conditional Random Field anyway?

- Random field because random variables are indexed in higher [than a single] dimensional space.
- Conditional because we can factorize relationships based on dependencies. And we can think of the factors in terms of being conditionally independent.

Linear-chain graphical model



Linear Chain CRFs

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right)$$

$$\mathbf{Z}(\mathbf{x}) = \sum_{\mathbf{y}} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x_t})\right)$$

Linear Chain CRF loss

$$l(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_t^{(i)}, y_{t-1}^{(i)}, \mathbf{x}_t^{(i)}) - \sum_{i} N \log Z(\mathbf{x}^{(i)}))$$

How does the training data look.

BIO Labelling Scheme

O	O	O	O	B	I	I	I	I	O	O
“We	really	needs	a	children	’	s	play	area	so	we

Features

- The identity of the word
- Is the word in a dictionary of entities
- Is the word capitalized.
- What is the part of speech of the word
- The shape of words in a window
- The shape of the current word
- Is the word capitalized
- Is in title form
- etc

Parameter Estimation & Inference

- Linear chain CRFs loss is convex w.r.t to θ_k.
- We use maximum likelihood and we maximize the the conditional log likelihood.
- Train using your favourite optimization technique: BFGS works, SGD works.
- Inferences of $p(y_t, y_{t-1}|x_t)$ done using the forward-backward algorithm.
- Prediction or the best sequence $\mathbf{y}* = \arg\max_{\mathbf{y}} P(y|x)$ is done using viterbi.

Packages

- crf++
- crfsuite
- Factorie
- Alchemy
- Pycrfsuite (from crfsuite)

What's the bottleneck?

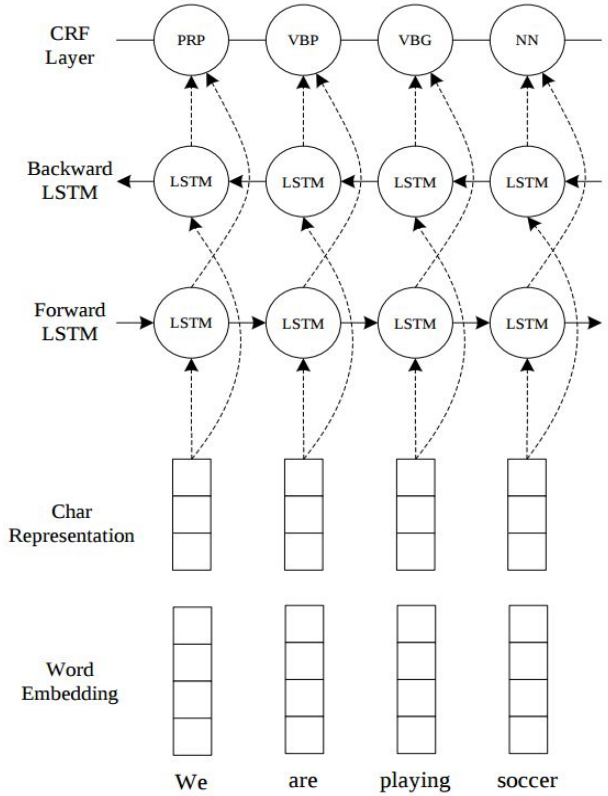
Hand-crafted features!

CNN-BI-LSTM-CRF

End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF from Ma and Hovy at Carnegie Mellon.

- 100D Glove Vectors
- Uniform initialization of weight, bias and Embedding vectors
- Uses a bag of tricks but nothing obscure: dropout, early stopping
- Achieves state of the art (91.7 F1 score) performance on the CoNLL-2003

Have a look


CRF Layer
PRP
VBP
VBG
NN
Backward LSTM
LSTM
LSTM
LSTM
LSTM
Forward LSTM
LSTM
LSTM
LSTM
LSTM
Char Representation
Word Embedding
We
are
playing
soccer

Let’s Build This Thing with Keras

https://colab.research.google.com/drive/1LunAisDJ8UyHHp1W6s-5uGQ8InsK46x2

Recap

- Named Entity Recognition (NER) is the task of determining entities (e.g phrases of interest like a New York Times, Obama, AUD)
- CRF's are powerful models to capture complex dependencies structures in data
- CRF can be combined with LSTMs to achieve end-to-end NER pipelines

Questions?

@orsonady on almost all the social things.