Effective Sequence Labeling with Hybrid Neural-CRF Models

Pablo Botton da Costa
Gustavo Henrique Paetzold

PROPOR - 2018

pablo.botton.costa@gmail.com
ghpaetzold@outlook.com

26 de Setembro de 2018

Summary I

Summary II

What is sequence labeling?

- Assigning sequences of labels to sequences of some objects is a very common task (NLP, bioinformatics);
- In general, learn a function h : $\sigma^* \rightarrow L^*$ to assign a sequence of labels from L to the sequence of input elements from σ.
- Sequence labeling tasks in NLP: Speech recognition, POS tagging, chunking (shallow parsing) and named-entity recognition.

Named Entity Recognition

- Named Entity Recognition is a process where an algorithm takes a string of text (sentence or paragraph) as input and identifies relevant nouns (people, places, and organizations) that are mentioned in that string;
- West Indian all-rounder **Phil Simons**_PERSON took four for 38 on Friday as Leicestershire...

Part Of Speech Tagging

- A part-of-speech (PoS) tagger is a tool that labels words as one of several categories to identify the word's function in a given language.
- Functions such as: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction and interjection.

What is meaning and how to represent it?

- "Meaning is use." Ludwig Wittgenstein;
  - The box concept
- "Contextual information is very important.- Paul Grice.
- How to represent meaning in a computational way?
  - Language models;
  - Distriburional semantics;
  - Neural Language models;
  - tree structure.

Deep learning

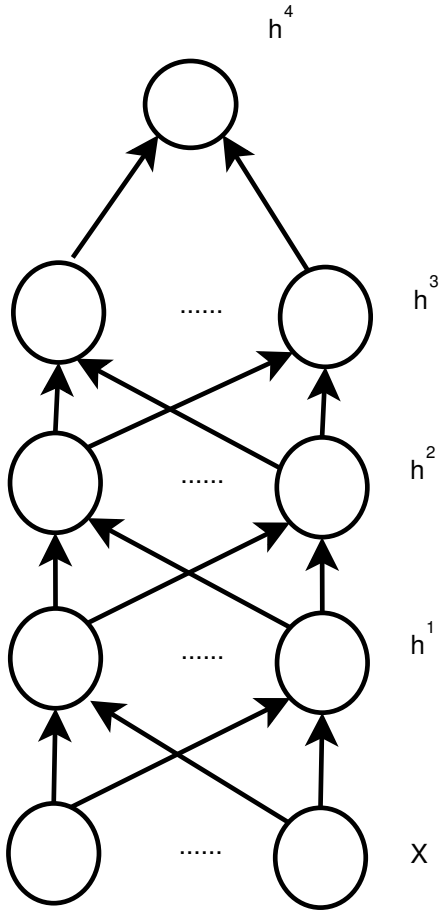


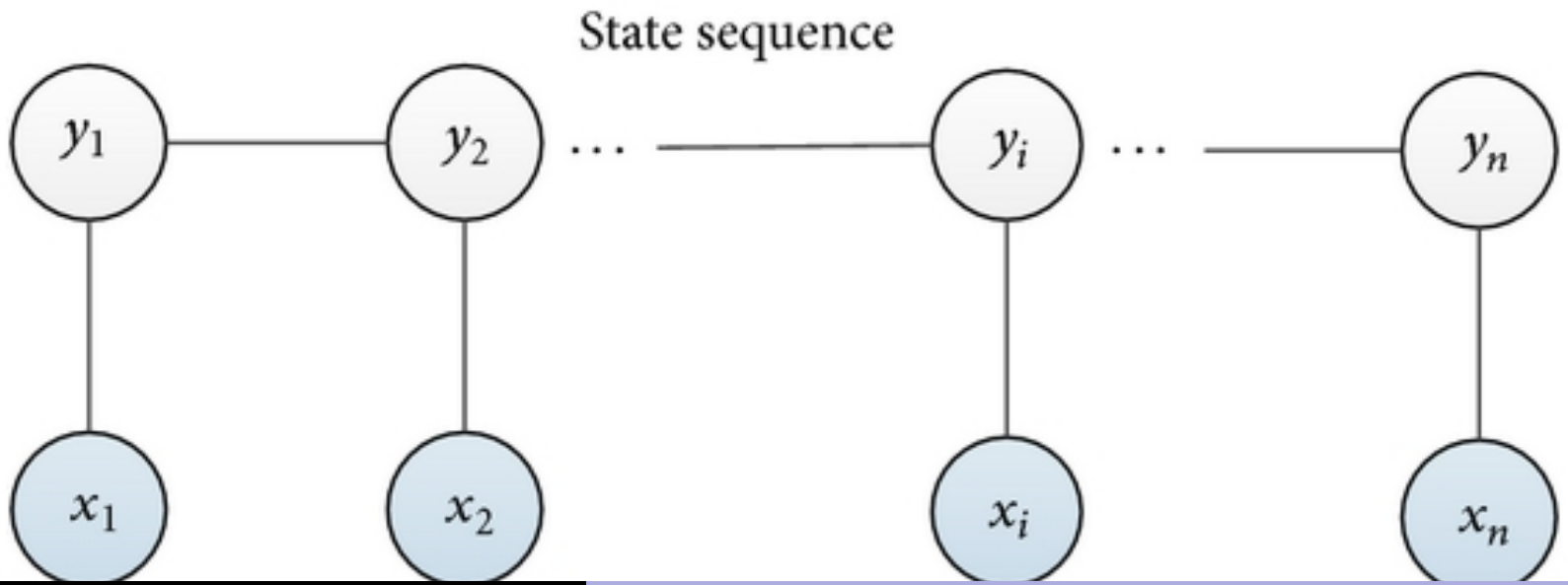
Figura: Adapted from [5]

Recurrent Neural Networks (RNN)

- What is it?
- Sequential feature induction. Very similar idea to a HMM.
  - Long dependencies, is it a problem? (The vanishing problem);
  - Long Short-Term Memory (long memory). Is it a solution? Depends of what you are trying to model in it...

Conditional Random Fields (CRF)

- What is it?
- A sequential classification model. Very similar to a Bayesian network.
- Can be used as a sequential learner.



LSTMs + CRF?

- General formula of CRF: $P(y_1, \ldots, y_m) = \frac{e^{C(y_1, \ldots, y_m)}}{Z}$ or $P(\tilde{y}) = \frac{e^{C(\tilde{y})}}{Z}$
- The new label is the defined by all the previous observations.
- We trained the model using cross-entropy: $-\log(P(\tilde{y}))$

Our goal

- Represent words and sentences in way that any model can extract features from it, and predict any kind of sequential textual task;
- While training a general model.

Data-sets

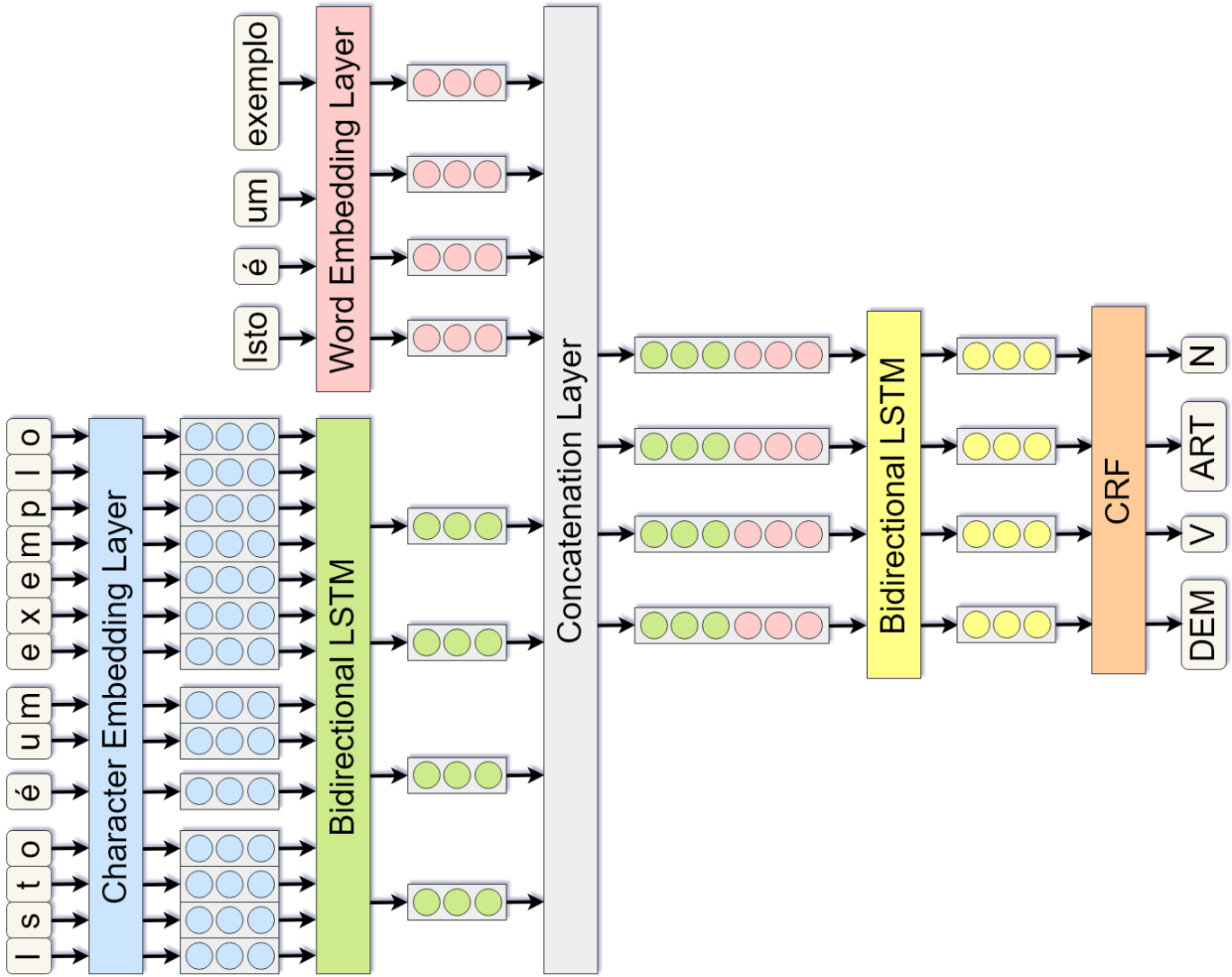
Tabela: Table defining the used corpus.

Corpus	Train		Test	
	Sentences	Tokens	Sentences	Tokens
HAREM I	4,749	93,125	3,393	62,914
UD 1.2	20,200	242,702	2,244	54,777

Word embeddings:

- At the final paper version we used the fasttext word embeddings.[1]

[1] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md


Isto
é
um
exemplo
Word Embedding Layer
I s t o
é
u m
e x e m p l o
Character Embedding Layer
Bidirectional LSTM
Concatenation Layer
Bidirectional LSTM
CRF
N
ART
V
DEM

Evaluation and scoring

$$P = \frac{\text{Total of classes} \cap \text{Number of correct ones}}{\text{Number of correct classes}} \quad (1)$$

$$R = \frac{\text{Total of classes} \cap \text{Number of correct ones}}{\text{Total of classes}} \quad (2)$$

$$WeightedF - score = \frac{2 \cdot P \cdot R}{P + R} \quad (3)$$

Results PoS Tagging

Tabela: Table comparision of our model results at universal dependencies corpus version 1.2 against state of art results.

Model	Accuracy	F-score	Precision	Recall
Full	97.87	97.58	97.52	97.64
Word+CRF	94.70	94.06	94.11	94.01
Word-NoCRF	94.70	94.10	94.20	94.00
[6]-Adversarial learning	98.07	-	-	-
[6]-NO$_A D$	97.94	-	-	-
[4]-BILSTM	97.90	-	-	-
[4]-TNT	96.27	-	-	-
[4]-CRF	96.32	-	-	-
[1]-Berend	95.50	-	-	-
[3]-Nguyen	97.50	-	-	-

Results NER

Tabela: Table comparing our model results at HAREM corpus against state of art results.

Modelo	Accuracy	F-score	Precision	Recall
Full	94.01	69.14	68.95	69.34
Word+CRF	91.67	60.56	66.83	55.36
Word-NoCRF	90.21	51.61	52.85	50.43
[2]Santos, et al.-CNN	-	65.41	67.16	63.74

Conclusions

- Superior or comparable to former state-of-the-art approaches;
- Incorporating both character-to-word embeddings and a CRF layer into the model yields considerably better results than using just typical word embeddings models;
- Our model offers state-of-the-art performance, requires no engineered features, and effectively handles rare and out-of-vocabulary words;
- Code[2].

[2] `https://bitbucket.org/pablocosta/deepnlptoolkit.git`

References I


Gábor Berend.
Sparse coding of neural word embeddings for multilingual sequence labeling.
arXiv preprint arXiv:1612.07130, 2016.


Cícero Nogueira dos Santos and Victor Guimarães.
Boosting named entity recognition with neural character embeddings.
CoRR, abs/1505.05008, 2015.

References II


Dat Quoc Nguyen, Mark Dras, and Mark Johnson.
A Novel Neural Network Model for Joint POS Tagging and Graph-based Dependency Parsing.
In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 134–142, 2017.


Barbara Plank, Anders Søgaard, and Yoav Goldberg.
Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss.
arXiv preprint arXiv:1604.05529, 2016.

References III


Richard Socher, Chris Manning, and Yoshua Bengio.
Deep learning for nlp (without magic).
In *ACL 2012 + NAACL 2013 Tutorial*, 2013.


Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev.
Robust multilingual part-of-speech tagging via adversarial training.
arXiv preprint arXiv:1711.04903, 2017.

Questions?

Effective Sequence Labeling with Hybrid Neural-CRF Models

Pablo Botton da Costa
Gustavo Henrique Paetzold

PROPOR - 2018

pablo.botton.costa@gmail.com
ghpaetzold@outlook.com

26 de Setembro de 2018