Title : Natural Language Processing (Almost) from Scratch
Writer : COLLOBERT, WESTON, BOTTOU, KARLEN, KAVUKCUOGLU AND KUKSA
Citations : 5359

Mohammad Sabik Irbaz (160041004)

January 22, 2020

Types of NLP task :

Four standard NLP tasks are described : Part-Of-Speech tagging (POS), chunking (CHUNK), Named Entity Recognition (NER) and Semantic Role Labeling (SRL).

Part-Of-Speech Tagging :

POS aims at labeling each word with a unique tag that indicates its syntactic role, for example, plural noun, adverb, . . .

Chunking or Shallow Parsing :

Aims at labeling segments of a sentence with syntactic constituents such as noun or verb phrases (NP or VP). Each word is assigned only one unique tag, often encoded as a begin-chunk (e.g., B-NP) or inside-chunk tag (e.g., I-NP). Validation is achieved by splitting the training set.

Named Entity Recognition :

NER labels atomic elements in the sentence into categories such as 'PERSON' or 'LOCATION'. As in the chunking task, each word is assigned a tag prefixed by an indicator of the beginning or the inside of an entity.

Semantic Role Labeling :

SRL aims at giving a semantic role to a syntactic constituent of a sentence. Feature categories commonly used by these system include (Gildea and Jurafsky, 2002; Pradhan et al., 2004):

- the parts of speech and syntactic labels of words and nodes in the tree;
- the node's position (left or right) in relation to the verb;
- the syntactic path to the verb in the parse tree;
- whether a node in the parse tree is part of a noun or verb phrase;
- the voice of the sentence: active or passive;
- the node's head word; and
- the verb sub-categorization.

Steps :

(Using Neural Network)

Transforming Words into Feature Vectors :

the first layer of their network maps each word indices into a feature vector, by a lookup table operation. Given a task of interest, a relevant representation of each word is then given by the corresponding lookup table feature vector, which is trained by backpropagation, starting from a random initialization.

Extracting Higher Level Features from Word Feature Vectors:

Feature vectors produced by the lookup table layer need to be combined in subsequent layers of the neural network to produce a tag decision for each word in the sentence. Producing tags for each element in variable length sequences is a standard problem in machine-learning. They considered two common approaches which tag one word at the time: a window approach, and a (convolutional) sentence approach.
Window Approach : A window approach assumes the tag of a word depends mainly on its neighboring words.
Sentence Approach : Window approach performs well for most natural language processing tasks we are interested in but this approach fails with SRL. In this case, tagging a word requires the consideration of the whole sentence. When using neural networks, the natural choice to tackle this problem becomes a convolutional approach. In the semantic role labeling case, this operation is performed for each word in the sentence, and for each verb in the sentence.
Tagging Schemes : In the window approach, these tags apply to the word located in the center of the window. In the (convolutional) sentence approach, these tags apply to the word designated by additional markers in the network input. POS task consists of marking the syntactic role of each word. However, the remaining three tasks associate labels with segments of a sentence.

Training :

They trained their neural networks by maximizing a likelihood over the training data, using stochastic gradient ascent. They used two ways for interpreting neural network outputs as probabilities.
WORD-LEVEL LOG-LIKELIHOOD : In this approach, each word in a sentence is considered independently. Using conditional probability and applying a softmax operation over all tags likelihood percentage is determined.
SENTENCE-LEVEL LOG-LIKELIHOOD : In tasks like chunking, NER or SRL we know that there are dependencies between word tags in a sentence: not only are tags organized in chunks, but some tags cannot follow other tags. Training using a word-

level approach discards this kind of labeling information. They considered a training scheme which takes into account the sentence structure: given the predictions of all tags by their network for all words in a sentence, and given a score for going from one tag to another tag, they want to encourage valid paths of tags during training, while discouraging all other paths.

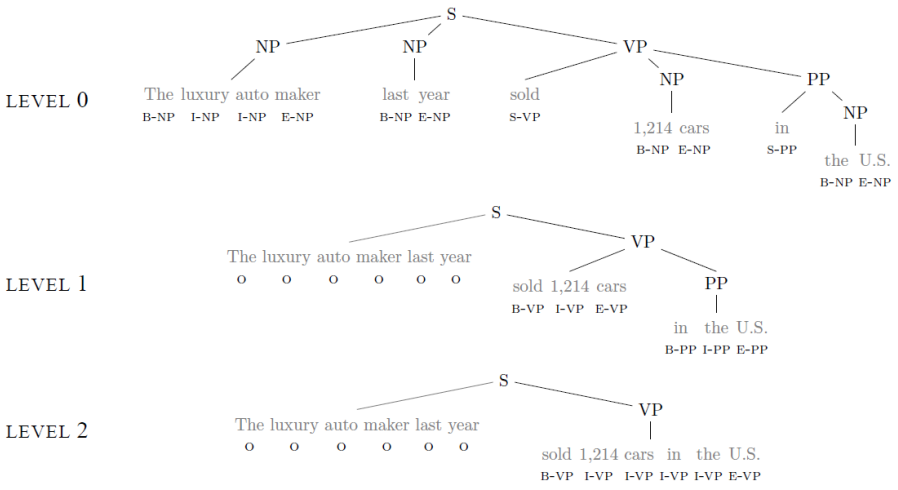

Figure 1: Parsing

Datasets :

Their first English corpus is the entire English Wikipedia. They used a dictionary containing the 100,000 most common words, with the same processing of capitals and numbers. Again, words outside the dictionary were replaced by the special 'RARE' word.