Syntactic Parsing Review

Zvengin
zvengin@nii.ac.jp

December 18, 2017

1 Syntactic Parsing

Syntactic parsing is the task of recognizing given sentence and assigning a syntactic structure to it. Syntactic parsing is an imaportant intermediate stage of representation for semantic parsing. The most sever problem in syntactic parsing is ambiguity called structural ambiguity. Structural ambiguity arises when the grammar can assign more than one parse to a sentence. There are two main kinds of structural ambiguity, one is attachment ambiguity which refers to that a particular constituent can be attched to a parse tree at more than one place. The other is coordination ambiguity, which indicates that different sets of phrases can be conjoined by a conjunction like 'and'.

1.1 CKY Parsing

CKY algorithm is dynamic programming. The adavantage of CKY algorithm arises from the context-free nature of grammar rules, that is, we store subtrees which have been parsed in table and we can directly use these parsed subtrees in subsequent parsing without reparsing them. The prerequisite of doing this is that grammar rules are context-free.
If there is a sentence S with length N. We construct a table T whose size is $(N+1) \times (N+1)$ and we just keep upper tri-angle matrix.
(1) Insert index bwtween two adjacent words in sentence.
for example: ${}_0\ a_1\ cat_2\ is_3\ running_4\ on_5\ the_6\ grass_7$
(2) Fill each word in corresopnding cell in table T.
for example: word "${}_3\ running_4$" should be filled in cell $T(3,4)$
(3) For column j, we calculate cell value of $T(i,j), i = 0, ..., j-1$ from bottom to up. With respect to cell $T(i,j)$, we select a k from $i+1, ..., j-1$ in order. With respect to a specific k, we check if $T(i,k)$ and $T(k,j)$ fit in with exsiting rules. If so, fill the right hand side of corresponding rule into cell $T(i,j)$. If not, check the next k. Therefore, there may exist several values in cell $T(i,j)$. Note that if $T(i,k)$ or $T(k,j)$ has several values, we need to check all possible combinations.
(4)Repeat (3) for each column from left to right.
(5)Choose parsing tree from cell $T(0,n)$ and then recursively retrieving its component from table

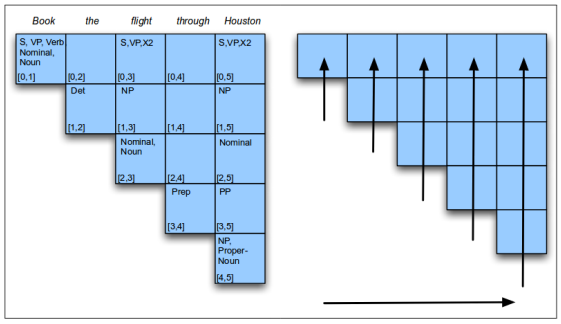

Figure 1: This figure shows the working principle of CKY

1.2 Partial Parsing

Partial parsing refers to that some applications doesn't require complete and complex parse tree, only partial parsing is sufficient.
(1)One method to tackle this problem is use cascade FSTs. cascade FSTs generates flatter parse tree which links all the major constitutes in an input.
(2)The other method is about chunking. Chunking is to identify and classify non-overlap segments of a sentence that constitute the basci non-recursive phrase corresponding to the major parts-of-speech found in most wide-converage grammars. This set typically includes noun phrases,verb phrases,adjective phrases, and prepositional phrases. Not all application require the identification of all of these categories.
(2.1)How to denote chunking in sentence.
We just use a bracketing notation to denote the location and type of a given sentence since chunked text lacks hierarchical structure.
(2.2)How to obtain chunked text
Currently we regard chunking problem as a sequence labelling problem which is similar to part-of-speech-tagging and use machine learning method to tackle this problem. We first define a IBO tagset and label each segment of sentence with target label in target set. For each type of chunk, it has unique start label B and intermediate label I and shared end lable o. Therefore, there are $2 \times n + 1$ tag in tagset.
The input feature is extracted from a context-window which surrounds the word to be classified. Using a window which extends 2 words before and 2 words after the word being classified. Feature extracted from this context-window include the words themselves, their part-of-speech, the chunk tags of the preceding inputs in the window. Therefore, a feature vector consists of 12 elements, if including the target of the word being classified, there will be 13 elements in input feature.
(2.3)How to obtain training data set
We extract our training data set from tree banks since it is very expensive to annotate each sentence by human.
(2.4)Evaluate Method
Similar to part-of-speech, we also compare the output of annotator with gold-standard answers provided by human annotators. However, we adopt precision,recall, and F-measure borrowed from information retrivel instead of using word-by-word accuracy measure used in part-of-speech tagging. This article refers[1]

References

[1] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London, 2014.