Tutorial 9: Deep-diving into Shallow Parsing

Alok Debnath

8 March 2019

1 Introduction

In the last tutorial, I covered the concept of multiword expressions in linguistics and NLP. Most methods of detecting and working with multiword expressions, named entity recognition and so on were based on dictionaries which had these phrases frozen. However, note that these dictionaries are inefficient to maintain over a long period in the general domain. On the other hand, you would note that most MWEs and Named Entities are contiguous in text, that is to say that a group of words which are syntactically close to one another are more likely to be related in such a manner.

Let us look at an entirely different problem in analytic languages such as Hindi. Hindi has a tendency to isolate markers which are non-essential to the meaning of the word it is being used with. This most commonly occurs with verbs, wherein Hindi has separate tense and aspect markers. Future perfect continuous constructions in English are another example of such clusters of words which are closely syntactically related to one another. This property is called Local Word Grouping and is considered a standard pattern in multiple languages. LWG relies on the fact that certain groups of words, which are close to one another, are a part of the same semantic unit. (This is where the concept of compositionality comes in).

In this tutorial, we look at chunking, also known as shallow parsing, which is a computational method of understanding and analyzing syntactically related units in text. under the assumption that they are also going to be semantically related. We shall also see the drawbacks of this assumption, how various languages and processing frameworks understand chunking, and some applications in downstream tasks.

2 Concept of Shallow Parsing

Phrase structure parsing is a common hierarchical representation of a sentence which shows the relation of one sequence of words to another sequence of words in the same sentence. Computationally, the boundaries between one group of words and another is arbitrary, in the sense that there is no fixed position or length of these related syntactic components.

Generative grammars are often used to overcome this with respect to phrase structure parsing in general. A generative grammar is a format of representing the hierarchical structure of language. An example of this is:

```
S → NP VP
NP → (Det) (AdjP)* NP | Noun
VP → (Adv)* VP (PP)* |  (Aux)* Verb
AdjP → (Adj)+
PP → Prep NP
Noun → cat | mat
Verb → sitting
Aux → was | had been | is | has been
Prep → on
Det → the | a
```

These are called generative grammars for their ability to generate sentences which are syntactically valid based on the phrase structure rules defined by the language. Penn Treebank has a generative phrase structure grammar of its own. Primarily, phrase structure parsing is recommended for languages with a fixed word order and therefore not very popular in Indian languages. PS trees also capture syntactic ambiguities, such as the PP attachment problem, among others, which are evolving fields of research today. PS parsing can be done in two ways, top down and bottom up. Bottom up parsing essentially starts from the words, POS tags them, and groups them together into their individual phrases, which are then grouped together again to form the entire sentence.

Chunking is called shallow parsing as it follows the procedure of phrase structure parsing bottom up, to a depth of two. This means that chunking is done by grouping syntactically related parts of speech together. Sometimes, chunks capture named entities, local word groups and even compounds, but this is not always a given. Interestingly, while PS tree parsing contains more information, multiple PS trees can exist for a single sentence. Moreover, PS trees are syntactically restraining, which is why we mentioned that it is preferred for languages with a fixed word order. However, chunking has no such restrictions, as local word groups, a form of chunking, is common for Indian languages.

3 Local Word Grouping in Indian Languages

Local word grouping is defined a bit more rigidly for Indian languages. Given that shallow parsing works under the assumption of semantic relation via syntactic closeness, word order movements based constituency tests are one of the most popular methods of identifying chunks. However, syntactic constraints make this a difficult process. One can define a strong local word group as one which has a strong internal structure, wherein permutations of this order are considered ungrammatical. Therefore, a weak word group would allow internal reconfiguration, which means that each strong local word group is a component of a weak word group.

The process of creating a local word group is very similar to the process of shallow parsing, except that in Indian Languages there is a need to do morphanalysis, POS tagging and chunking. One of the major questions now, is the order in which this is to be done. Think about it.

4 Useful References

1. Abney, Steven P. "Parsing by chunks." Principle-based parsing. Springer, Dordrecht, 1991. 257-278. : http://www.vinartus.net/spa/90e.pdf

2. Jurafsky, Dan, and James H. Martin. Speech and language processing. Vol. 3. London: Pearson, 2014. Section 13.5 25-31

3. Bharati, Akshar, Vineet Chaitanya, and Rajeev Sangal. "Local word grouping and its relevance to Indian languages." Frontiers in Knowledge Based Computing (KBCS90), VP Bhatkar and KM Rege (eds.), Narosa Publishing House, New Delhi (1991): 277-296.: https://faculty.iiit.ac.in/~sangal/web/files/papers/kbcs90.pdf