Hybrid of Rule Based and Probabilistic Parsing

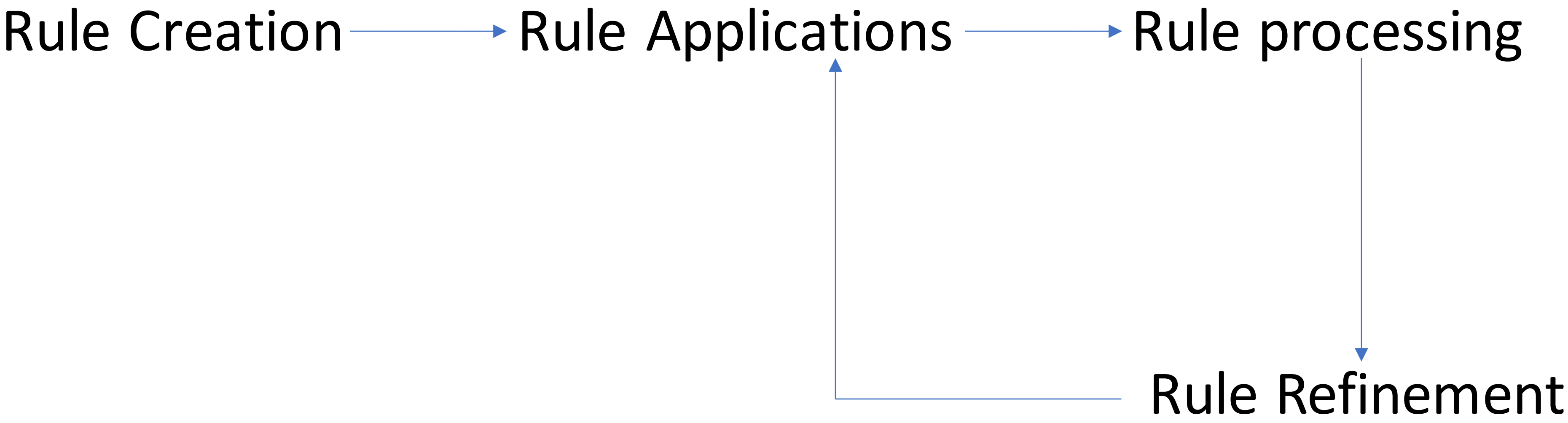
Rule Based Parsing

- Rule based parsing Approach is one of the oldest NLP Methods in which predefined linguistics rules are used to analyze and process textual data.
- Rule based Approach involves Applying a particular set of rules or patterns to capture specific structures, extract information, or perform tasks such as text classification and so on.
- A Parser that uses Hand Written(Designed) rules as opposed to rules that are derived from the data.

Steps in Rule Based Approach in NLP

- **Rule Creation** : Based on the desired tasks, domain specific linguistics rules are created such as grammar rules, syntax patterns, semantics rules or Regular expressions.
- **Rule Application :** The Predefined rules are Applied to the inputted data to capture matched patterns.
- **Rule Processing :** The text data is processed in accordance with the results of the matched rules to extract information, make decision or other tasks.
- **Rule Refinement:** The created rules are iteratively refined by repetitive processing to improve accuracy and performance.

Steps in Rule Based Approach



Rule Creation

Rule Applications

Rule processing

Rule Refinement

Steps in Rule Based Approach

Advantages of Rule Based Approach

- Easily Interpretable as rules are explicitly defined
- Rule-based techniques can help semi-automatically annotate some data in domains where you don’t have annotated data (for example, NER(Named Entity Recognition) tasks in a particular domain).
- Functions even with scant or poor training data
- Computation time is fast and it offers high precision
- Many times, deterministic solutions to various issues, such as tokenization, sentence breaking, or morphology, can be achieved through rules (at least in some languages).

Disadvantages of Rule Based Approach

- Labor-intensive as more rules are needed to generalize
- Generating rules for complex tasks is time-consuming
- Needs regular maintenance
- May not perform well in handling variations and exceptions in language usage
- May not have a high recall metric

Why Rule-based Approach with NLP

- Rule-based NLP usually deals with edge cases when included with other approaches.
- It helps to speed up the data annotation. For instance, a rule-based technique is used for URL formats, date formats, etc., and a machine learning approach can be used to determine the position of text in a pdf file (including numerical data).
- Also, in languages other than English annotated data is really scarce even for common tasks which are carried out by Rule-based NLP.
- By using a rule-based approach, the computation performance of the pipeline is also improved.

Probabilistic Parsing

- Probabilistic parsing is using dynamic programming algorithms to compute the most likely parse(s) of a given sentence, given a statistical model of the syntactic structure of a language.
- The parser uses a set of rules and probabilities in probabilistic parsing to determine the most likely syntactic structure for a sentence.
- This kind of structure allows the parser to take into account the context of the sentence and the likelihood of different syntactic structures and select the most probable parse given the context.

Probabilistic Model

- Similar to a Context Free Grammar, a probabilistic context-free grammar G can be defined by a *G*=(*M*,*T*,*R*,*S*,*P*) where:
- M is the set of non-terminal symbols
- T is the set of terminal symbols
- R is the set of production rules
- S is the start symbol
- P is the set of probabilities on production rules

Probabilistic Lexicalized CFG

- Probabilistic Lexicalized Context-Free Grammar (PLCFG) is also a type of grammar used in natural language processing to generate and analyze sentences in a given language.
- It is a combination of a lexicalized context-free grammar which uses lexical items that word as the basic units for generating sentences, and probabilistic models which assign probabilities to the different rules and structures in the grammar.
- Probabilistic Lexicalized Context Free Grammar solve the somewhat separable weaknesses that stem from the independence assumptions of PCFGs, out of which the most often remarked on one is their lack of lexicalization.

- https://www.google.com/search?q=rule+based+parsing+video&oq=&gs_lcrp=EgZjaHJvbWUqCQgAECMYJxjqAjIJCAAQIxgnGOoCMgkIARAjGCcY6gIyCQgCECMYJxjqAjIJCAMQIxgnGOoCMgkIBBAjGCcY6gIyCQgFECMYJxjqAjIJCAYQIxgnGOoCMgkIBxAuGCcY6gLSAQozNTk1NzhqMGo3qAIIsAIB&sourceid=chrome&ie=UTF-8#fpstate=ive&vld=cid:0015f78d,vid:uPLVGQfptYk,st:0