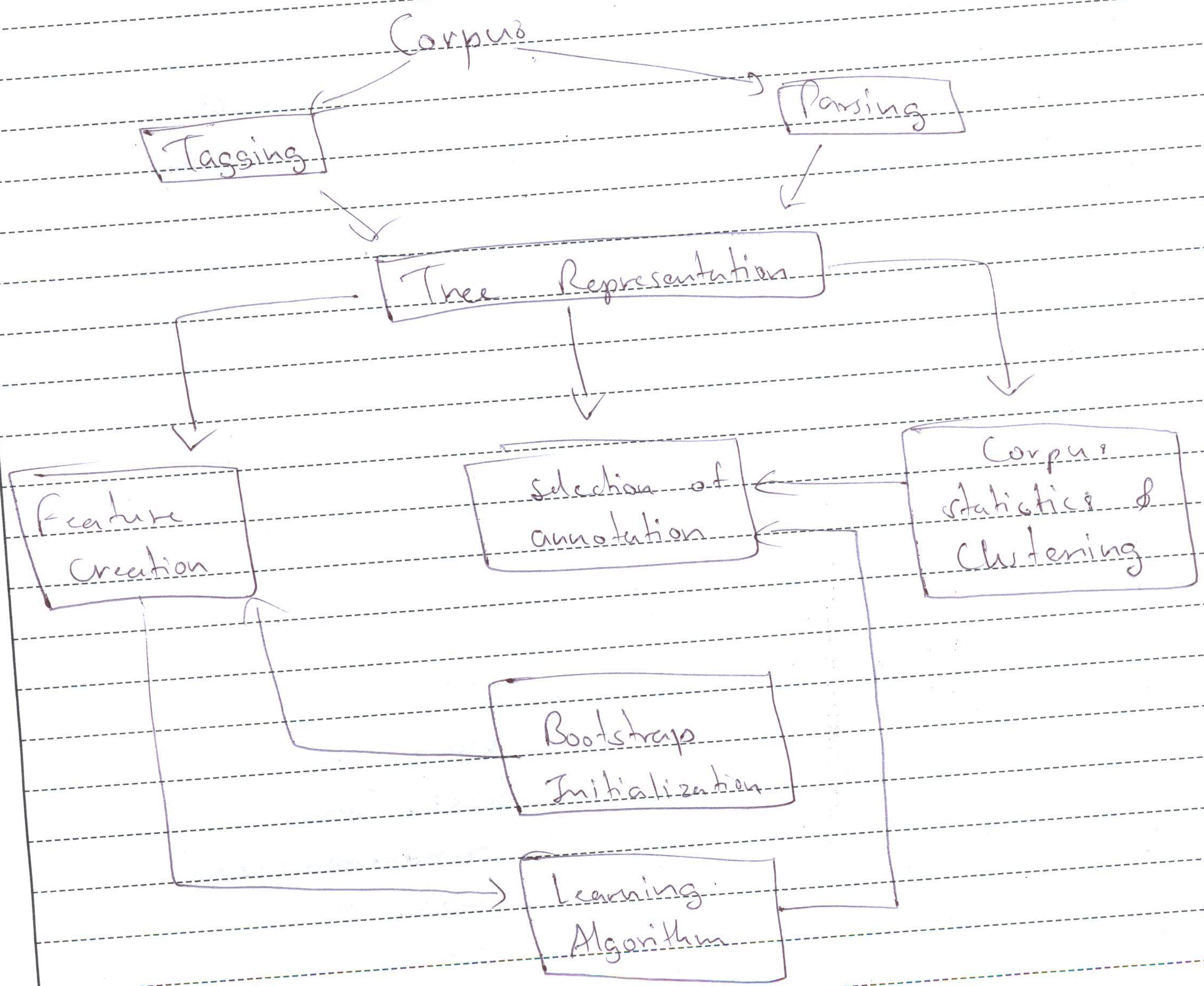
Iat - 3

1) With a neat diagram explain the architecture used in the task of learning to annotate cases with knowledge roles.



* Collection of 500 microsoft Document written in Germany amounting to one million words
* Transform the document into XML format
* Extract paragraph belonging to cases.
* Perform POS tagging
* Perform syntactical parsing

* Transform results into XML representation for manual annotation
* Construct features for learning algorithm
* Implement an active learning strategy

Input Data:
- Raw text: The input data consists of raw text documents or sentences that need to be annotated with knowledge roles.
- Annotations: Existing annotations that serves as the training data.

Text Processing:
- Tokenization: The text is broken down into smaller units, usually words or subwords
- Embeddings: Convert tokens into numerical vectors using techniques like word embeddings

Knowledge Role Annotation Model:
- Neural Network: Utilize a neural network architecture suitable for sequencia labeling tasks. Common choices include Bidirectional LSTMs, GRUs, or transformers (BERT).
- Pre-trained Models: Pre-train the model on a large corpus if possible, especially if using transformer-based models.
- Fine-tuning: Fine-tune the model on your specific task using annotated data.

Loss function:
- Define a loss function suitable for sequence labeling tasks, such as cross-Entropy Loss.

Training:
• Train the model on the annotated dataset, adjusting the model parameters to minimize the defined loss function

Evaluation:
• Assess the model's performance using a separate evaluation dataset, common metrics include precision, recall, and F1 score.

Inference: • Deploy the trained model to annotate new, unseen text data.

Post-Processing:
• Optional: Apply post-processing techniques to refine annotations, depending on the specific requirements of your task.

2) Give detailed description of the following
a. Domain knowledge
b. Domain Concept
c. knowledge Roles
d. Global security.org

a) Domain knowledge:
• Domain knowledge refers to the backround knowlege about a particular industry or field of study.
• It includes concepts, entities, relationships, facts etc. relevant to that domain.
• For example, in the medical domain, knowledge about disease, medications, procedures is considered domain knowledge.

b) Domain Concept:

- A domain concept is a key idea or topic of interest within a domain.
- For example, in the finance domain, stocks, bonds, IPO's are important concepts.
- Identifying and understanding domain concepts is vital for domain-specific NLP.

c) Knowledge Roles:
- Knowledge roles refer to the types of relationships b/w entities expressed in text.
- For example, an attack event may have knowledge roles like perpetrator, victim, instrument etc.
- Annotating text spans with knowledge roles aids in info extraction and knowledge base construction.

d) Globalsecurity.org:

- Globalsecurity.org is a public policy organization that provides news and analysis about security threats and military affairs.
- It covers global security topics like military operations, cybersecurity, terrorism etc.
- The site aggregates news from multiple sources and provides original commentary and analysis.
- It serves as a data source for researching global security events and risks.

3. Explain the strategies used in the active learning approach ~~the~~ a for acquiring labels using a committee-based classification scheme.

The active learning approach with committee-based classification is a strategy that leverages a committee of diverse learners to optimize the process of acquiring labeled data. Here's a breakdown of the key strategies involved:

Maintain a Committee of Learners:

• Create a committee of diverse learners. This diversity can be achieved by training each learner using different algorithms or by using different initializations.

Predict Labels for Unlabeled Examples:

• For instances that lack lablels, predict their labels using two random committe members. This introduces variability into the predictions.

Query Selective Sampling:

• If the predictions from the committe members disagree on the label for an unlabeled example, consider this example as potentially informative. Request the true label for this example to resolve the disagreement

Update all Committe Members:

• Once a labelled example is acquired, use it to update all committee members. This ensures that the committee learns from the new information.

CMR

Committee Disagreement Signals Informativeness:

- Instances where the committee members disagree on predictions are considered informative. Disagreement signals uncertainity or complexity in the data.

Leverage Disagreement to Reduce Labeled Examples Needed:

- Actively seeks instances where the committee members disagree as resolving these instances tends to provide the most informative labels.
  This selective quering helps reduce the total no of labeled examples needed.

Use Majority Vote for final predictions:

- Combine the predictions of all committee members using a majority vote. This final prediction is likely to be more ~~tha~~ accurate than that of any single learner.

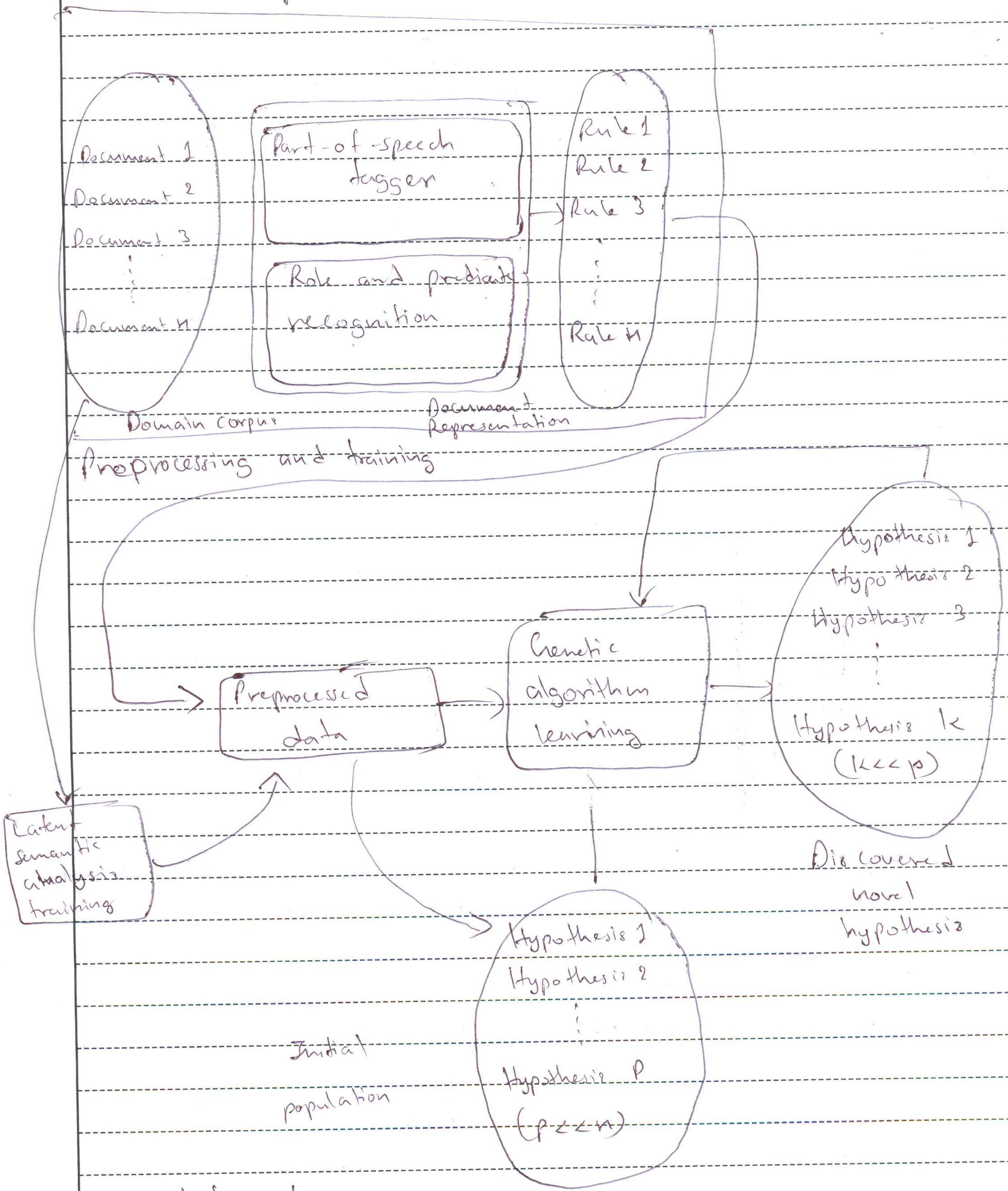
Committe Prediction Improves Over Time:

- As more labeled data is acquired and the committe is updated, the prediction accuracy of the committe as a whole improves. The diversity within the committe contributes to a more robust learning process.

4) With a neat diagram. Explain the evolutionary model for KDT (knowledge Discovery from Text)

Knowledge discovery: aims at producing & evaluating explanatory unseen hypothesis
Processing starts by performing IE tasks which applies extraction pattern and then Generates a rule

like representation for each document of specific domain corpus.



knowledge discovery

~~Ha~~ * After processing 'n' documents → produce 'n' rules
the rule should represent the document conten w.r.to
conditions & conclusion.
- Now, rules along with other training data will be
modeled to guide the fun based discovery
Rule → Training data
- To generate an initial set of hypothesis, an initial
population is created → from the initial rules
- GA runs for No. of generations until the fixed no
of generation is achieved → at last small set of
best hypothesis are obtained

IAT - 2

2b Explain Dependency - path kernel for relation extraction

A: All relevant words from a shortest path b/w the two
entities in a graph structure where edges correspond
to relations b/w a word (head) and its dependents.

* Arguments are connected to their target predicates
either directly through an arc pointing to the predicate
or indirectly through a preposition or infinitive particle

• [Local Dependencies] These correspond to local predicate-
argument constructions such as troops → raided,
or 'pumping → stations'

• [Non-local Dependencies] Long - distance dependencies arise
due to various linguistic constructions such as coordination,
extraction, raising and control. Among non-local
dependencies are ~~to~~ troops → warning or 'ministers →
preaching'.