SLIDE
Research Methodology in I.T.
Lecture 08 - A Template-based Approach to Design an Experiment

Author: Dr. Rao Muhammad Adeel Nawab
Instructor: Dr. Rao Muhammad Adeel Nawab

SLIDE
Research Cycle – Three Main Phases
1. **Reading**
2. **Writing**
3. **Experimentation**

SLIDE
Lecture Outline
- **Introduction**
- **Research Focus and Expected Outcome(s)**
- **Main Components of an Experiment**
- **Template-based Approach to Design an Experiment**

SLIDE
==========
Introduction
==========

SLIDE
Experimental Design
- **Definition**
  - **A process of planning and designing, to clearly understand how to run / conduct an experiment (or set of experiments)**
- **Purpose**
  - **To ensure that set of experiment(s) are practically possible and are likely to reveal expected outcome(s)**
- **Importance**
  - **It is important to properly plan and design an experiment (or set of experiments) before running / conducting it to ensure that**
    - **Expected challenges in running / conducted experiments are addressable**
    - **Appropriate resources (particularly data and computational resources) are available**

        - **Experiment will be run using standard practices in that field of research**
        - **Expected outcome(s) are likely to be achieved after running the experiment**
- **Applications**
    - **Properly planning and designing an experiment (or set of experiments)**
        - **Saves time and effort**
        - **Saves computational resources**
        - **Increases productivity and efficiency**
        - **Increases the chances of getting your desired outcomes from experiment**

SLIDE
==========
Research Focus and Expected Outcome(s)
==========

SLIDE
Research Focus

- **Problem**
    1. **Paraphrased Text Reuse Detection**
- **Baseline Approach**
    1. **Exact Matching Approach**
        - **An existing approach**
- **Proposed Approach**
    1. **Query Expansion Approach**
        - **You have proposed this approach to further improve the performance for the paraphrased text reuse detection problem**

SLIDE
Expected Outcome(s) of Experiment

- **Proposed approach will outperform the baseline approach for the same problem on the same dataset / corpus**

SLIDE
Expected Outcome(s) of Experiment (Cont…)

- **Very Important**
- **If you can't "theoretically prove" that your proposed approach will outperform the baseline approach, then it will be very risky to carry out the experiment(s)**

SLIDE

Expected Outcome(s) of Experiment (Cont…)

- **Example – Baseline Approach vs Proposed Approach**
  - **Problem: Paraphrased Text Reuse Detection**
  - **Dataset**
    - **Original Text: A dog bites a man**
    - **Reused Text: A hound bites a person**
  - **Evaluation Methodology**
    - **Quantify degree of similarity between a text pair using the following formula**
    - **Similarity Score = No. of Common Words / Total Number of Words in Reused Text1**
  - **Exact Matching Approach (Baseline Approach)**
    - **Original Text: A dog bites a man**
    - **Reused Text: A hound bites a person**
    - **Similarity Score = No. of Common Words / Total Number of Words in Reused Text = 3 / 5 = 0.6**
  - **Query Expansion Approach (Proposed Approach)**
    - **Original Text: A dog bites a man**
    - **Reused Text: A hound (dog) bites (cuts) a person (human, man)**
    - **Similarity Score = No. of Common Words / Total Number of Words in Reused Text = 5 / 5 = 1.0**
  - **Original Text and Reused Text are semantically same; therefore, similarity score should be 1.0**
  - **Similarity Scores, Approach (HINT: MAKE A TABLE)**
    - **0.6 - Exact Matching Approach (Baseline Approach)**
    - **1.0 - Query Expansion Approach (Proposed Approach)**
  - **Conclusion**
    - **Proposed Approach outperforms the Baseline Approach for a sample dataset**
    - **Based on this finding it is likely that Proposed Approach will outperform the Baseline Approach on large dataset**

SLIDE
==========
Main Components of an Experiment
==========

SLIDE
Main Components of an Experiment

- **Techniques / Approaches / Methods**
- **Dataset(s) / Corpora**
- **Evaluation Methodology**
- **Evaluation Measures**

SLIDE
Techniques / Approaches / Methods

- **Two Options**
    1. **Write Detailed Description of Techniques the Experimental Setup Section**
    2. **Just Mention Names of Techniques**
        - **Refer to previous Section(s) which contain detailed description of techniques**

SLIDE
Techniques / Approaches / Methods

- **Detailed Description of Technique**
- **For each technique describe the following main points**
    1. **In which category it falls?**
    2. **How it works?**
    3. **How you have used it in your study?**
    4. **In which previous studies it has been used?**

SLIDE
Dataset(s) / Corpora

- **For each dataset / corpus mention**
    1. **Name of Corpus**
    2. **Main Characteristics of Corpus**

SLIDE
Evaluation Methodology

- **Important – Always use standard tools and methods mentioned in high quality research papers**
- **Main Points**

- **Problem**
- **How Problem is Treated**
- **Tools / Toolkits**
  - **How tools / toolkits were used in your study**

- **Example**
  - **Problem – Paraphrased Text Reuse Detection**
  - **Treated as – Supervised Text Classification Task**
  - **Toolkit - WEKA**

SLIDE
Evaluation Measures

- **Use standard evaluation measures**
  - **HINT: See what evaluation measures are used in previous studies which addressed the same problem**
- **For each evaluation measure**
  - **Write Definition**
  - **Write Equation**

==========
Template-based Approach to Design an Experiment
==========

SLIDE
Make a Template to Design an Experiment

- **Make a Template**
  - **Use Divide and Conquer Approach**
    - **Step 1: Break "Experimental Setup Section" into Sub-sections / Sub-sub-sections**
    - **Step 2: For each Sub-section / Sub-sub-section**
      - **Summarize each paragraph into "one single sentence"**
    - **Step 3: From each Paragraph**
      - **Extract important information in the form of "Attribute-Value Pair"**
  - **Order and Flow**
    - **A high-quality Experimental Setup Section must tell a connected and coherent story**
    - **Understand the Order and Flow in Experimental Setup Section at three levels**
      - **Sub-section / Sub-sub-section Level**
        - **Connection between Paragraphs**
      - **Paragraph Level**

- **Connection between Sentences**
- **Sentence Level**
  - **Connection between Words / Phrases**

- **Independence and Connectivity**
  - **In a high-quality Experimental Setup Section, each**
    - **Sub-section**
    - **Sub-sub-section**
    - **Paragraph**
    - **Sentence**
  - **Must be both independent and connected**

SLIDE

Steps - Template-based Approach to Design an Experiment

- **Step 1: Shortlist 2 – 5 research papers very closely related to your research work**
- **Step 2: For each paper**
  - **Extract information from existing shortlisted papers in the form of "Attribute-Value Pair"**
  - **Fill the template (discussed in previous slides)**
- **Steps 3: Combine the experimental designs of existing papers to make your template-based experimental design**
- **Step 4: Discuss your experimental design with your Supervisor / Senior for further suggestions and refinements**
- **Step 5: Repeat Steps 1 to 4, until you and your supervisor are happy with the experimental design**
- **Step 6: Starting running / conducting experiments ☺**

SLIDE

Example - Template-based Approach to Design an Experiment

- **Paper Title - Detecting Text Reuse with Modified and Weighted N-grams**
- **Authors - Rao Muhammad Adeel Nawab, Mark Stevenson, Paul Clough**
- **Conference - *SEM 2012: The First Joint Conference on Lexical and Computational Semantics**

SLIDE

Example - Template-based Approach to Design an Experiment

- **Template – Experimental Design**
  1. **Dataset**
     - **METER Corpus**

- **Original (or Source) Documents = 771**
- **Reused (or Derived) Documents = 768**
  - **Wholly Derived = 285**
  - **Partially Derived = 300**
  - **Non-Derived = 183**

- **Pre-processing**
  - **Lower case**
  - **Remove all punctuation marks**

2. **Approaches**
   - **N-gram Overlap Approach (Baseline Approach)**
     - **Word N-grams (N = 1 - 5)**
     - **Containment Similarity Measure**
       - **Similarity Score Range is [0 – 1]**
   - **Modified and Weighted N-grams Approach (Proposed Approach)**
     - **Modified N-grams**
       - **Deletion Approach**
       - **Substitution Approach**
         - **WordNet**
         - **Paraphrase Lexicon**
     - **Weighted N-grams**
       - **Weight of N-gram = Probability of N-gram (-log (P))**
       - **Probability of N-gram Computed using Language Model**
       - **Language Model**
         - **Trained on Reuters Corpus**
         - **Using SRILM Language Modeling Toolkit**
     - **Word N-grams (N = 1 - 5)**
     - **Containment Similarity Measure (with Clipped Count)**
       - **Similarity Score Range is [0 – 1]**
3. **Evaluation Methodology**
   - **Problem = Paraphrased Text Reused Detection**
   - **Solution = Treated as Supervised Document Classification Task**
   - **Two Versions of Classification**
     - **Ternary Classification – Differentiate between 3 Classes**
       - **Wholly Derived**

- **Partially Derived**
- **Non-Derived**

- **Binary Classification – Differentiate between 2 Classes**
  - **Derived (Wholly Derived + Partially Derived)**
  - **Non-Derived**

- **Machine Learning Toolkit**
  - **WEKA 3.8**
  - **Implemented in Java Programming Language**
- **Classifiers / Machine Learning Algorithms**
  - **Naïve Bayes**
- **Classifier Training / Testing**
  - **K-fold cross validation approach (K = 10)**
- **Features used as input to Naïve Bayes**
  - **Similarity scores generated for Word N-grams (N = 1 - 5) using**
    - **N-gram Overlap Approach**
    - **Weighted and Modified N-gram Approach**

4. **Evaluation Measures**
   - **Three Evaluation Measures used for evaluation**
     - **Precision**
     - **Recall**
     - **F_1**
   - **Approaches performance measured using**
     - **Macro-averaged Precision, Recall and F_1 scores across all classes**

TODO: NEED TO PUT THIS EXAMPLE EXPERIEMNTAL DESING IN THE EXCEL SHEET

SLIDE
Your Turn

- **Read the following paper Experimental Setup Section and fill the template (in Excel Sheet) mentioned in the lecture**
  - **Note that you can add / subtract "Attributes" from the template discussed in the lecture**
- **Paper Title - Comparing Medline Citations using Modified N-grams**
- **Authors - Rao Muhammad Adeel Nawab, Mark Stevenson, Paul Clough**

- **Journal – JAMIA**

SLIDE

Summary - A Template-based Approach to Design an Experiment

- **Experimental design can be defines as a process of planning and designing, to clearly understand how to run / conduct an experiment (or set of experiments)**
- **The main goal of an experimental design is To ensure that set of experiment(s) are practically possible and are likely to reveal expected outcome(s)**
- **Main benefits of an experimental design are**
  - **Saves time and effort**
  - **Saves computational resources**
  - **Increases productivity and efficiency**
  - **Increases the chances of getting your desired outcomes from experiment**
- **The main Expected Outcome of an Experiment is that**
  - **Proposed approach will outperform the baseline approach for the same problem on the same dataset / corpus**
  - **Very Important - If you can't "theoretically prove" that your proposed approach will outperform the baseline approach, then it will be very risky to carry out the experiment(s)**
- **Main Components of an Experiment**
  1. **Techniques / Approaches / Methods**
  2. **Dataset(s) / Corpora**
  3. **Evaluation Methodology**
  4. **Evaluation Measures**
- **How to Make a Template for an Experiment**
  - **Use Divide and Conquer Approach**
    - **Step 1: Break "Experimental Setup Section" into Sub-sections / Sub-sub-sections**
    - **Step 2: For each Sub-section / Sub-sub-section**
      - **Summarize each paragraph into "one single sentence"**
    - **Step 3: From each Paragraph**
      - **Extract important information in the form of "Attribute-Value Pair"**
  - **Order and Flow**
    - **A high-quality Experimental Setup Section must tell a connected and coherent story**

- **Understand the Order and Flow in Experimental Setup Section at three levels**
  - **Sub-section / Sub-sub-section Level**
    - **Connection between Paragraphs**
  - **Paragraph Level**
    - **Connection between Sentences**
  - **Sentence Level**
    - **Connection between Words / Phrases**
- **Independence and Connectivity**
  - **In a high-quality Experimental Setup Section, each**
    - **Sub-section**
    - **Sub-sub-section**
    - **Paragraph**
    - **Sentence**
  - **Must be both independent and connected**

- **Steps - Template-based Approach to Design an Experiment**
  - **Step 1: Shortlist 2 – 5 research papers very closely related to your research work**
  - **Step 2: For each paper**
    - **Extract information from existing shortlisted papers in the form of "Attribute-Value Pair"**
    - **Fill the template (discussed in previous slides)**
  - **Steps 3: Combine the experimental designs of existing papers to make your template-based experimental design**
  - **Step 4: Discuss your experimental design with your Supervisor / Senior for further suggestions and refinements**
  - **Step 5: Repeat Steps 1 to 4, until you and your supervisor are happy with the experimental design**
  - **Step 6: Starting running / conducting experiments ☺**