Generating Machine Readable Experimental Procedures from Papers

6.864: Menghsuan Sam, William Spitzer 6.806: Iveel Tsogsuren, Renqiao Zhang

Experimental procedures in scientific papers are human (at least the expert in the field) readable, but usually difficult for computer due to various reason. Missing or implied information as well as complicated actionables often arose due to the concise nature of scientific papers and for protecting intellectual properties of the research group. In the age of automation, this problem limits the ability of automatic experiment conducting robots. As a result, this project aims to translate procedure into machine readable instruction. To reduce the scope of the project, our group has decided to focus on electrochemical related paper, but we want to point out that with the same set of resource, our model can be generalize to any field of research.

Our model will treat the experimental procedures as recipes. A procedures detail a series of instructions to convert ingredients into a result with pre-specified steps just like a recipe. In addition, experimental procedures must be straightforward so that results can be emulated by different teams. We are interested in applying machine learning and natural language processing techniques that can take in a recipe and convert the procedures into a standardized machine readable instructions. Our recipe corpus will be experimental procedures from scientific journals for electrochemical societies.

We approach this problem by performing the following tasks:

Building a database of common procedure frameworks
Our framework database will be built using the textbook, Electrochemical Methods: Fundamentals and Applications by Bard and Faulkner, consisting of common electrochemical experimental procedures. The framework will consist of a set of ingredients and steps that must be done to complete the experiment. The framework will also detail which ingredients and steps are required, and which ones are replaceable. This task requires NLP or manual framework tagging of experimental procedures from the textbook in a machine readable form, then creating a database that allows for easy Matching between ingredients/actions and experiment.

Training a POS language model to identify the ingredients and actions from procedures from papers.
We require a Parts of Speech tagging model to identify words in the experimental procedures as ingredients or actions. This model will use the tagging techniques that we've learned in class (e.g. n-grams, Viterbi, Maxent) to produce a set of ingredients and steps for each procedure. This tasks requires building a POS tagging model with high accuracy that can extract ingredients and actions from a procedure.

Match ingredients and actions to the best framework from our database
We need a method that finds the framework using a set of ingredients and actions from our database. This task requires a scoring method that compares the set of ingredients/actions from

the procedure in the paper against the set from each framework in the database. The method then picks the framework with the best score to use.

Fill in the experimental framework from our database using tagged ingredients and actions

Using the framework from the database, we need to fill in the relevant ingredient and actions using the information that we've extracted from the procedure. This task also includes figuring out a method of filling in missing information, or modifications to the framework based on the procedure.

Evaluation Method for Procedure Correctness

We will test our algorithm against procedures found in papers from various Scientific Journals (refer to Available Resources). Since our algorithm produces a unique machine readable format, we will need to manually evaluate our output. We will check 2 parameters: accuracy on deciding the framework model, and accuracy in producing an output that matches the ingredients and steps found in the original procedure text.

Evaluation Method for Machine Readability

We will build a simple script to evaluation machine readability and executability. For example, is the procedure too complicated, or include all relevant information? Does the machine have all the required ingredients and equipment?

Available Resources:

Electrochemical Methods: Fundamentals and Applications by Bard and Faulkner
Advanced Functional Materials
Electrochemistry Communication
Langmuir
Journal of Electrochemical Society
Journal of Solid State Electrochemistry
Electrocatalysis

Separation of Tasks:

Deciding on Framework Structure - William + Menghsuan
Building the framework database, ingredients/actions database - Renqiao + Iveel
POS tagging model to identify ingredients and actions - Iveel + William
Max Entropy matching between ingredients and actions, and framework - Renqiao + Menghsuan
Filling in Framework with relevant ingredients/actions. (Includes solving the missing ingredient/actions problem) - William + Menghsuan
Evaluation Correctness - Menghsuan + All
Evaluation Machine Readability - William