AMS-MIDAS Institution Disambiguation Project Report

Yile Gu, Arun Massand, Lokraj Srinivasan, Jonathan Gryak

April 2021

1 Introduction

Mathematical Reviews (MR) is a division of the American Mathematical Society. Since 1940, MR has been collecting data on research literature in mathematics. From the beginning, institutional affiliations of authors have been collected, primarily to distinguish authors with similar names. However, institutional identities are themselves susceptible to ambiguities in their naming, especially since organizations are identified down to the level of departments, not just universities or colleges. Under the current convention for institution codes, the codes follow the pattern A-BB-CCC, where A is one or more character for the country, BB is two or more characters that represent a university or other large organization, and CCC represents the unit at the level of a department. The main objective of this project is flagging possible duplicate or ambiguous institution codes within the given 207,334 institutions in the rows of data.

2 Methods

A schematic diagram of the proposed method for institution disambiguation is depicted in Figure 1 below.

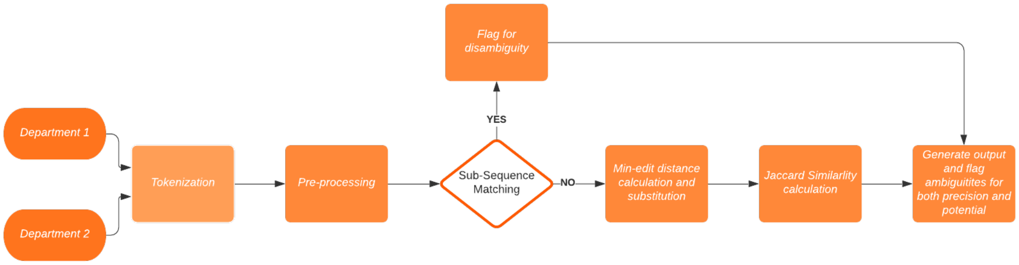
Flag for disambiguity
Department 1
Department 2
Tokenization
Pre-processing
YES
Sub-Sequence Matching
NO
Min-edit distance calculation and substitution
Jaccard Similarlity calculation
Generate output and flag ambiguitites for both precision and potential

Figure 1: Schematic overview of disambiguation process

The method begins by first tokenizing and removing special characters from department names. Next, stop-words, such as “of” and “the” as well as common department words, such as “department” and “institute”, are filtered from the tokenized department names. The stop-words as well as common department words are removed from the department names in order to prioritize words that actually define department names when calculating similarity scores. For example, a department named “The Department of Electrical Engineering and Computer Science” would become [“Electrical”, “Engineering”, “Computer”, “Science”] after tokenization and filtering.

After preprocessing, every pairwise combination of departments is created within the given university. Next, sub-sequence comparisons are conducted between the tokenized department names to find perfect sub-strings between names (e.g.,“Department of Math” and “Department of Mathematics”) which are then flagged as potential disambiguations. Figure 2 below demonstrates a simple example between “Department of Computer Science” and “Department of Computer Science and Engineering”.

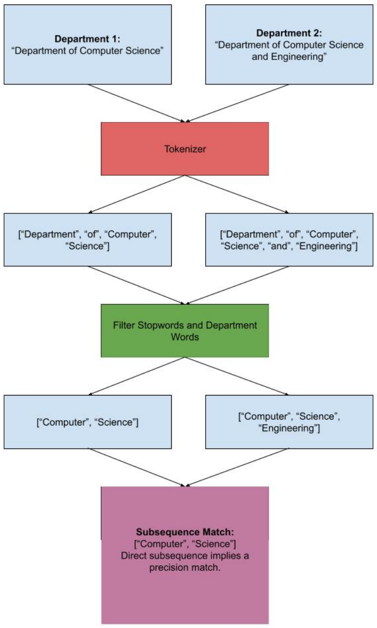
Department 1:
“Department of Computer Science”
Department 2:
“Department of Computer Science and Engineering”
Tokenizer
[“Department”, “of”, “Computer”, “Science”]
[“Department”, “of”, “Computer”, “Science”, “and”, “Engineering”]
Filter Stopwords and Department Words
[“Computer”, “Science”]
[“Computer”, “Science”, “Engineering”]
Subsequence Match:
[“Computer”, “Science”]
Direct subsequence implies a precision match.

Figure 2: Sub-sequence detection between “Department of Computer Science” and “Department of Computer Science and Engineering” within Korean University

If the two given department names are not a perfect sub-string match, the Damerau-Levenshtein minimum edit distance method is then used to compute possible pairs of words that might be spelling mistakes or close matches. The Damerau–Levenshtein distance between two words is the minimum number of operations (i.e., insertions) required to change one word into the other. If a certain word in a department's name is within an edit distance range of four insertions to another word in the second department's name, those words are then considered a matched word. The set of one department's name is then altered to contain the same matched word. After the minimum edit distance alterations, a Jaccard Similarity metric is calculated on the tokenized department names, yielding a similarity coefficient score between the two names. The Jaccard Similarity Coefficient is a statistic used for gauging the similarity between the two sets of words, and is defined as the size of the intersection divided by the size of the union of the two sets:

$$\text{jaccard}(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

If that score is above the defined threshold, the two departments are then considered to be similar enough and could be a potential disambiguation. In order to find the optimal values for both the Jaccard Similarity Coefficient threshold as well as the Damerau-Levenshtein minimum edit distance threshold, experimentation and tuning were conducted using the two parameters on a geographically diverse set of universities - Korean University, Vilnius University, University of Melbourne, State University of Campinas (UNICAMP), and University of Nigeria. After tuning the parameters using these training universities, final testing and analysis was conducted using four institutions - Princeton University, Georgia Institute of Technology, Yale University, and University of Split.

An example disambiguation from Korean University is provided in Figure 3.

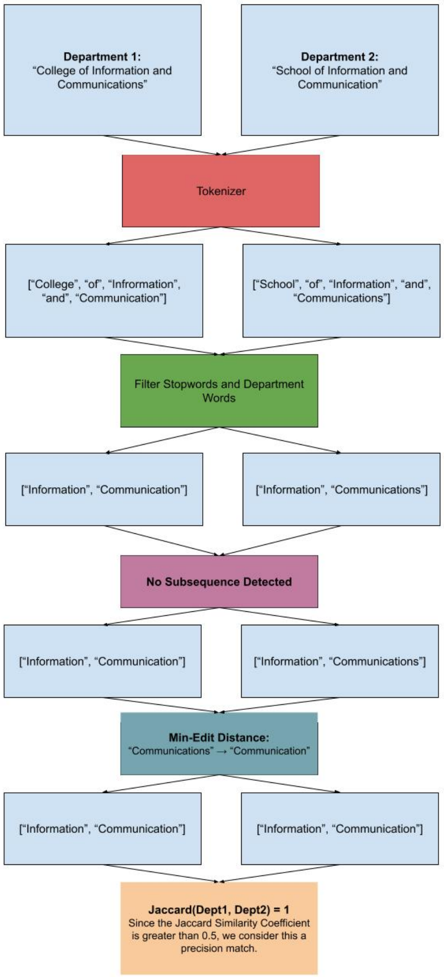
Department 1:
"College of Information and Communications"
Department 2:
"School of Information and Communication"
Tokenizer
["College", "of", "Infrormation", "and", "Communication"]
["School", "of", "Information", "and", "Communications"]
Filter Stopwords and Department Words
["Information", "Communication"]
["Information", "Communications"]
No Subsequence Detected
["Information", "Communication"]
["Information", "Communications"]
Min-Edit Distance:
"Communications" → "Communication"
["Information", "Communication"]
["Information", "Communication"]
Jaccard(Dept1, Dept2) = 1
Since the Jaccard Similarity Coefficient is greater than 0.5, we consider this a precision match.

Figure 3: An example disambiguation detection for the institutions "College of Information and Communications" and "School of Information and Communication" within Korean University.

3 Experimental Results

The results of the disambiguation process are structured around the idea of Precision versus Potential duplicate department matches. By utilizing two different sets of parameters the model is able to achieve results with varying degrees of confidence. The first set of parameters, the Precision set, has slightly higher and more scrutinizing threshold values for the minimum Jaccard Similarity Coefficient and the maximum edit-distance value. The second set of parameters, the Potential set, has slightly less restrictive values for these same values. The parameters chosen for the precision matching was a Jaccard similarity of greater than 0.5 and less than or equal to 2 edits, while the potential parameters were 0.5 or greater and less than or equal 4 edits. The parameter range for the Jaccard similarity ranged from 0.35 to 0.65 in 0.05 intervals and the minimum edit distance was between 2 and 6 edits. Using these sets of parameters, the method aims to create more comprehensive and useful results. Subsequently, the results are structured as follows for a given department: the school name, the department name, the current 7-digit institution code, followed by a list of all the precision-based parameter matches within the same university along with a list of those matches department codes and a list of all the potential-based parameter matches within the same university and their department codes as well. This output was performed for each department for all universities within the dataset. A selected set of rows from Korean University is displayed below. Only one pair for each mode is included in Table 1 for brevity.

Precision	Precision Codes	Potential	Potential Codes
'Institute of Statistics' , 'Department of Statistics'	'KR-KOR-IS', 'KR-KOR-S'	'Institute of Economics', 'Institute of Economic Research'	'KR-KOR-IEC', 'KR-KOR-ECR'

Table 1: Example results from Korean University

As can be observed in Table 1, 'Institute of Statistics' and 'Department of Statistics' are a Precision match. This is because after removing stop-words and common department words such as 'Institute', 'Department' and 'of', the two department names are an exact match. In terms of Potential match, the calculation for 'Institute of Economics' and 'Institute of Economic Research' is slightly more comprehensive. After removing stop-words and common department words, the min-edit distance method will detect that 'Economics' and 'Economic' are two words that are sufficiently close and substitute one for the other. This will boost the Jaccard similarity above the chosen threshold, making it a Potential match. Table 2 below is a summary of the training universities mentioned earlier. These schools were chosen in particular due to their varied geographic and cultural properties, which allowed for better analysis into how the method performed for institutions from different countries and areas. Table 4 that follows is a summary of the testing universities discussed previously.

University Name	**Department (n)**	**Precision**	**Potential**
Korean University	113	98.63%	92.57%
Vilnius University	52	96.15%	94.92%
University of Melbourne	130	98.43%	68.12%
State University of Campinas	128	96.36%	88.51%
University of Nigeria	12	100.0%	100.0%

Table 2: Training universities with their associated number of departments along with precision and potential accuracies.

Method	**Mean Accuracy (SD)**
Precision	97.914% (1.63%)
Potential	88.824% (12.29%)

Table 3: Mean accuracy and standard deviation (SD) of the precision and potential methods on the training set.

University Name	**Department (n)**	**Precision**	**Potential**
Princeton University	87	96.66%	75.00%
Georgia Institute of Technology	110	98.71%	80.46%
Yale University	119	97.43%	58.65%
University of Split	41	96.96%	100.0%

Table 4: Test universities with their associated number of departments along with precision and potential accuracies.

4 Limitations

Upon completion of testing and manual inspection of the results, while the proposed method works fairly well, it is not perfect. For example, there was a 96.66% accuracy rate for disambiguating departments of Princeton University using the Precision parameters and 75.00% accuracy rate with the Potential parameters. The reason for misclassifications was not because of duplicate departments that were missed within Princeton University, but instead due to unique cases of two differing department names being considered syntactically similar (small edit distances with subsequent high Jaccard score) but not in

Method	**Mean Accuracy (SD)**
Precision	97.44 % (0.90%)
Potential	78.53% (17.05%)

Table 5: Mean accuracy and standard deviation (SD) of the precision and potential methods on the test set.

meaning. This is handled by the fact that output is separated into Precision versus Potential matches. In almost all cases, these “misclassifications” are due to instances where the minimum edit distance threshold allows for large variance in word changes and thus a subsequently large jaccard similarity that primarily occurs with the Potential match column.

A limitation of the current program is that it doesn’t necessarily work optimally for department names that are not in the English language. The stop-words and department words that are filtered out are all in the English language, so if an institution’s department is listed in any language other than English, the stop-words and department words will not be removed, so the Jaccard Similarity Coefficient will be incorrectly calculated. Further, departments listed in different languages require different minimum edit distance parameters in order to work best. In order to optimize performance in different languages, further experimentation must be conducted with regards to how various languages’ special characters are treated. Depending on the language and the inclusion/exclusion of extra characters, an edit distance of 4 characters might not cover enough edit changes or could result in too many edit changes within a pairs of words, leading to poor results such as in the case of State University of Campinas (UNICAMP). For instance one of the output possibilities was between “Departamento de Física” and “Center for Logic, Epistemology and History of Science (CLE)”, which are clearly two separate departments but after preprocessing, end up being labeled as potential matches. In summation, the current implementation works well within English-based department names but future work could revolve around modifying the algorithm to accommodate different languages without necessarily requiring translation.

5 Conclusion

In this project, various data processing techniques have been utilized to detect duplicate department names and corresponding codes in institution information provided by the American Mathematical Society. The method utilizes word tokenization, subsequence detection, min-edit distance and Jaccard similarity to affect disambiguation. To provide different levels of confidence, two different sets of parameters are used - Precision and Potential - that utilize different thresholds for detecting duplicates. Precision mode is optimized for accuracy, while Potential mode provides users with more possible matches for further investigation. Future improvements on this project include adding new parameters or modifying the method to improve performance for department names written in languages other than English.