**Data Normalization**

This document is used to introduce the details of data normalization.

Data normalization is based on the types of values:

(1) Entity data, such as disease name, test name in relational database. When processing these values, we will match the entities with the knowledge base using similarity function. The similarity function σ considers the Levenshtein distance, trigram, and longest common substring, here is formula:

We use Area Under Curve (AUC) to optimize the threshold with a small-scale dataset with manual annotations. We also mention the step on page 7 of our paper. The performance for entity data normalization is divided into three categories, the concrete AUC scores are shown in the following table. Final, we apply the optimal thresholds to all entities, and compute the link rate which indicates the percentage of all entities that have been successfully linked to the knowledge graph.

|  |  |  |  |
| --- | --- | --- | --- |
| Entity type | Optimal threshold | AUC | Link rate |
| Disease | 0.40 | 0.76 | 0.80 |
| Drug | 0.40 | 0.85 | 0.79 |
| Assay | 0.70 | 0.82 | 0.64 |

(2) Non-entity data, such as test result and unit. These data in EMR have quality problems which include missing value, non-standard value, and so on. Give an example, the unit of test results is not the same in the identical test. For instance, the “blood glucose” result values use the units of mg/L and g/L. We need a uniform unit, and the result values are converted along with it. In addition, the following situations may occur: If a patient is diagnosed with multiple diseases, the name of the disease is recorded in a cell separated by punctuation and spaces. We need to separate them to make sure that each row in the record stores only one disease. We use the Google refine tool to process and unify these situations.