Combination of Rule-based and Textual Similarity Approaches to Match Financial Entities

Ahmad Samiei Ioannis Koumarelas Michael Loster Felix Naumann

Hasso Plattner Institute, Potsdam, Germany
{first-name.last-name}@hpi.de

1. INTRODUCTION

Record linkage is a well studied problem [1] with many years of publication history. Nevertheless, there are many challenges remaining to be addressed, such as the topic addressed by FEIII Challenge 2016[1]. Matching financial entities (FEs) is important for many private and governmental organizations. In this paper we describe the problem of matching such FEs across three datasets: FFIEC, LEI and SEC. We were able to achieve an f-measure of 93.78% in the first task, which is comparable to the maximum 97.44%, and 70.44% for the second task, where the maximum is 88.38%.

2. COMBINING RULES AND SIMILARITY FUNCTION

Our approach follows the intuition of combining domain specific rules and similarity-based matching functions to address the problem. However, before being able to do any of these steps we had to first clean the data. The most important steps we took are: *data cleansing*, *data enrichment*, and finally *record linkage*.

2.1 Data cleansing

Initially, we had to find common attributes that we could use for record linkage, which in most cases were the same as those that the organizers suggested in the guidelines. Consequently, besides other actions, we performed stemming, capitalization, special characters and redundant white space removal, *normalization, fixing inconsistent attributes and data extraction*. We briefly explain the last three steps:

In normalization the goal is to use a single word for all different representations of the same concept. For instance: "PO BOX", "Post Box Office", "BOX", "P.O. Box" were all transformed to "pobox". Moreover, we transformed *synonyms* (e.g., {"ROAD", "RD", "HighWay", "Avenue"} to "ST").

In some of the given attributes, we had to classify them or even split them, in accordance with the information they should represent. For instance, in the case of the address, a single attribute was provided, with combined information as the address and care-for information (provided with "C/O"). Even more specific information like suite, floor etc, would be given in separate attributes, but with no consistency, which needed further cleansing in order to be used.

Concerning the FE's name, we split the name of the entity into root and modifier. Consider the separation of "FIRST CLOVER LEAF BANK" and "NATIONAL ASSOCIATION", which are root and modifier respectively.

2.2 Semantic extraction

Based on the information we derived from the ground truth, it seems that the most determinant part of the entities names' in many cases is ambiguous. More specifically, the tokens that specify if a FE is a *bank*, *holding* or a *subdivision* of an institute are not always that obvious.

Marking an entity as a bank. In FFIEC, all of the records were FEs by definition. However, in the other two datasets, LEI and SEC, we had to find out which of the records refer to FEs, to focus on only relevant cases. By using term frequency (TF) of the tokens in FFIEC, we found out which tokens are the most representative for FEs. By *filtering* for records that contain one of the following tokens: bank, trust, savings, loan, national, federal, state, we managed to cover more than 99% of FFIEC entities. Thus, while filtering the LEI and SEC datasets, we expect that we could limit our input record set to FEs only. When applied to LEI and SEC the selectivities were 20% and 15%, respectively. We assume that this could help us for both efficiency (fewer unnecessary pair comparisons) and effectiveness (only match between FEs).

Inclusion of a descriptive token. From the provided ground truth, in the case of the SEC dataset, we had to take care of specific tokens referred to modulos, that if present change the outcome of the matching entirely. For instance: "TA", "MSD", "GFN", "ADR" and "BD" are some of the most commonly encountered[2]. In such cases, even if the addresses were different, we would consider them a match, since they still might be referrring to different departments of the same entity.

Marking as holding. In accordance with the ground truth, it was evident that there was a difference between a holding corporation of a bank and the bank itself. Unfortunately, finding out whether an entity is a holding or not, turned out to be a difficult task using only the given information. For instance, in some cases they were even providing the

[1] https://ir.nist.gov/dsfin/

[2] For the full list consider: https://www.sec.gov/forms

DSMM'16, June 26-July 01 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2951894.2951905

same address. Consider the example of a **bank** versus a **banc**. A banc (provided also with different forms: "bank corp", "banc", "bancorp" etc.) usually refers to a corporation, which in most cases represented the holding of the respective bank. These cases should not be considered a match despite their similarity. For instance, the ground truth provided the two terms "united community bank" and "united community bancorp", which refer to a bank and its respective holding corporation.

2.3 Data enrichment

For cases where the provided information was ambiguous, we tried to enrich our data with external resources. Ambiguity usually comes from acronyms or different sub-divisions that refer to the same entity, or more importantly lack of concrete evidence of whether a FE is in fact a "holding" or not. Our attempts to enrich the datasets are the following:

Business profile enrichment. Consulting web-sites like Google finance, Reuters, and Bloomberg, turned out to be a good way of finding extra information about a specific FE. Using the acquired information, we were able to resolve some cases, such as whether an FE was a bank or not. Unfortunately, there were some drawbacks in this enrichment process:

- No result: for most of the banks, we did not get any result back. For instance, community banks tend to be small and operate locally.
- Useless: the returned result was referring to some other bank. E.g., for "WASHINGTON TRUST BANK" we got "Zions Bancorporation".
- Misleading: the returned result was textually similar to the query record, but contradicting w.r.t. the semantics we defined in the Section 2.2. Consider the example: "BLUE HILLS BANK" and the result "Blue Hills Bancorp Inc", which as we described before, is a holding of the queried record.

In cases of correct match (less than 2% for the different datasets) with this extra information we could resolve ambiguities for the queried bank. Regarding the problems, further filtering results, could be one step or even using a larger combination of sources.

SWIFT codes. SWIFT or BIC codes uniquely identify FEs. There were two ways to use lists[3] of such codes, either as a reference knowledge base or its bank code subpart as a blocking key. Matching against these bank names was problematic, because of the different formats of names used for the same banks. This required three record linkage tasks instead of two, FFIEC⋈SWIFT, SWIFT⋈SEC and SWIFT⋈LEI, which in practice reduced our effectiveness.

The reason why none of the aforementioned solutions was chosen, was the fact that they seemed to propagate more errors in the end.

2.4 Record linkage

In order to link the records across the given datasets we combined two different record matching classifiers. Our first classifier relied solely on methods that determine textual similarities, whereas the second applied several rules to further prune the results. By combining both classifiers we were able to produce our best results.

[3]http://www.theswiftcodes.com/

	Precision	Recall	F-score
Task 1 - Solution 1	96.58%	91.13%	93.78%
Task 1 - Solution 2	60.30%	96.17%	74.13%
Task 1 - Max Achieved	99.24%	96.37%	97.44%
Task 2 - Solution 1	71.75%	69.57%	70.64%
Task 2 - Solution 2	41.74%	83.48%	55.65%
Task 2 - Max Achieved	92.82%	85.65%	88.38%

Table 1: Scoring of our approaches, compared to the maximum achieved in the contest.

Textual classifier. This classifier used the Damerau Levenshtein distance to compare cities and a customized version of Monge-Elkan for the FEs' names. This version in contrast to the original one, performs two alternative steps. First, it finds the most similar tokens between the two strings and does not consider them again. Second, this customized version follows the commutative property, so that the order of the strings is not affecting the results. This means that we repeat the process by swapping s1 and s2, and then return the mean of the results.

Ruled-based classifier. As there were many records containing very similar values with slight differences in their attributes, namely address and entity name, utilizing only a threshold-based approach seemed to be insufficient for classification. This approach combines thresholds and rules to match records more precisely. It separately computes similarity for different parts of location, such as address number, po box, suite, floor, street address, i.e., address without general terms. It finds the match and non-match addresses based on combination of thresholds and rules. In case of entity names it uses the same approach and computes similarities for their roots, modifiers, domain of operations (national, federal and state), descriptive tokens and the type of entities (bank, banc, backcorp, inc, corp, etc.). Again, the rules make a decision for matching entities' names and finally records themselves.

3. EVALUATION AND CONCLUSION

The two solutions, that we uploaded for the two main required tasks, aimed for precision and recall respectively, which is also reflected in the received results from the organizers (Table 1). In total, in Task 1 we achieved *f-measure* of 93.78%, which is close to the maximum achieved 97.44%, and for Task 2 our best f-measure was 70.64%.

Overall, the most decisive factor to devise an appropriate similarity measure, seems to be the domain knowledge, especially for such a complex domain. In the future, based on our failed attempts, we would like to focus on the challenges of enriching data with a combination of appropriate external resources. This enrichment should enhance our similarity measure, so that we can distinguish between matches and non-matches easier and more accurately.

4. REFERENCES

[1] P. Christen. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, 2012.