The Extremes of Good and Evil

Master Thesis

presented by
Earl Hickey
Matriculation Number 9083894

submitted to the
Data and Web Science Group
Prof. Dr. Right Name Here
University of Mannheim

August 2014

Contents

List of Figures

List of Tables

Chapter 1

Phase I - Data Translation

1.1 Use Case & Data Profiling

The goal of the project is to aggregate company information from several sources. To this end we used the suggestions from the project into slides as a foundation for our use case. We also included an additional source and amended the dbpedia query to extract further relevant information. Thus, the relevent entity will be a company. We relied on 4 different datasources which are profiled in Table 1.1. In order to being able to process the kaggle dataset, we had to filter it down. To this end, we used the ”size range” attribute and only kept the categories ”10001+”, ”5001 - 10000”, ”1001 - 5000”, ”501 - 1000”, ”201 - 500”, ”51 - 200”. This is a valid approach since the forbes dataset contains data about the 2000 largest companies in the world, and the dataworld (dw) dataset is also called ”largest companies”, containing only companies of a certain size.

1.2 Consolidated Schema & Transformations

The consolidated schema was created by hand. The following transformations were applied to the input datasets:

1. Monetary values were normalized to the same base.

2. The *dbpedia* dataset came with currency information. The intention was to use this information via a mapping table to convert all monetary values to USD. However, it turend out that there was a huge amount of currencies involved and it was not clear of which data the exchange rate should be retrieved. Therefore the currency was kept as an additional attribute.

3. A unique ID was generated for each record that was mapped to the target schema.

4. ...

Dataset	Source	Format	#E[1]	#A [2]	List Of Attributes
kaggle	link to dataset	csv	7.1M (491.830[3])	11	ID, name, domain, year founded (MV), industry, size range, locality (MV), country (MV), linkedin url, current employee estimate, total employee estimate
forbes	link to dataset	csv	2.000	9	Company, Country, Sales, Profits, Assets, Market Value, Sector, Industry
dataworld (dw)	link to dataset	csv	1.924	10	Global Rank, Company, Sales, Profits, Assets, Market Value, Country, Continent, Latitude, Longitude
dbpedia	Query provided in Appendix	json	3.986	11	Name, industry_label, domain, founding_year, ceos, no_emp (MV), country (MV), location (MV), revenue (MV), income (MV), assets (MV)

Table 1.1: Dataset Overview. All dataset only refer to the class "Company", * For hyperlinks pls refer,[1]# of Entities, [2]# of Attributes, [3] Number of filtered companies from the original dataset might be smaller than this number, because the final XML was extracted from the previous XML by mathching the filtered names (company with the same name might exist also in excluded category).

Class Name	Attribute Name	Attribute Type	Contained in DS...
Company	name	String	Kaggle, Forbes, dbpedia, dw
Company	domain	String	Kaggle, dbpedia
Company	Year founded	Integer	Kaggle, dbpedia
Company	Industry	String/List	Kaggle, Forbes, dbpedia
Company	Size_range	Category	Kaggle (can also be derived for dbpedia)
Company	locality	String	Kaggle, dbpedia
Company	Country	String	Kaggle, Forbes, dw, dbpedia
Company	Linkedin url	String	Kaggle
Company	Current employee estimate	Integer	Kaggle, dbpedia
Company	Total employee estimate	String	Kaggle
Company	Sales	Integer	Forbes, dw, dbpedia
Company	Profits	Integer	Forbes, dw, dbpedia
Company	Assets	Integer	Forbes, dw, dbpedia
Company	Market Value	Integer	Forbes, dw, dbpedia
Company	sector	String	Forbes
Company	Global Rank	Integer	Dw
Company	Latitude	Decimal	dw
Company	Longitude	Decimal	Dw
Company	ceos	list	Dbpedia

Table 1.2: Attribute Mapping

Chapter 2

Phase II - Identity Resolution

2.1 Gold Standard

In order to create the gold standard we ran initial identity resolutions with two cheap and a more complex matching rule. With a threshold of 0.2 we used three different matching rules:

1. Jaccard-3-Grams on company names (with frequent tokens removed) (MR1)
2. Levensthein Similarity on company names (with frequent tokens removed) (MR2)
3. A combination of 1, 2, Longest Common Subsequence, and a token-based similarity (MR27)

The results were then combined into one file which for each individual correspondence outlined the similarity calculated by every similarity measure as well as the company names. Furthermore, an average of the three similarities was calculated. We then labeled the matches into matching record pairs, corner-case matches and non-matches, and non-matches. Therefore, the correspondences were sorted by the average similarity. A high value indicated that all matching rules consider this correspondence a match. Correspondences were labeld as certain matches if the similarity scores had a high matching threshold > 0.9 and an actual match was present. Non-matches in this area were labeled as corner cases. For $0.9 > avg.sim. > 0.7$ mathces were labeled as corner-cases. Afterwards the correspondences with the lowest avg. sim. were reviewed and labeled as non-matches or corner-case matches. Then, the correspondences were in turn sorted by each individual similarity measure score and non-categorized correspondences were labeld as corner-cases above a threshold of 0.7. To achieve the distribution according to the rule of thumb

outlined in the lecture, which states to include 20% matching record matching pairs, 30% corner-case matches and non-matches (fuzzy), and 50% non-matching record pairs a random sample out of the labled correspondences was drawn. The data was then split into train and test set using a python script, with a test size of 0.25 and stratified on the gold standard category.

After running several identity resolutions the evaluation logs were analyzed and the gold standard was amended to cater to the identified false positive correspondences (according to the previous gold standard) by adding them to the gold standard, usually as corner-cases.

2.2 Matching Rules

2.2.1 General Setup

The company name was the sole variable we could rely on during the identity resolution. To this end, we implemented several string-based comparators which are outlined in table 2.2.1 and subsequently combined them to form different matching rules. We implemented a plethora of different matching rules, therefore we limit ourselves to the top 10 based on their F1-scores. Every comparator had options to include certain pre- and post-processing steps. Preprocessing steps generally included lowercasing, removal of punctuation, and removal of whitespaces (the latter was omitted for token based similarity metrics). We also implemented post-processing capabilities. These included a threshold after which the similarity was set to zero, and an option to boost or penalize the similarity based on a certain threshold. For example, a similarity might be boosted up using a particular function above a threshold of 0.8 and penalized below.

			Preprocessing		
ID	Similarity measure	Parameters	PWL[1]	Rm FT[2]	Focus
1	Jaccard on ngrams	n: ngram length	✓	(✓)	Overlap
2	Jaccard on tokens		*	(✓)	Overlap
3	Levensthein		✓	(✓)	Typos / Edit-distance
4	Longest Common Subsequence	Normalization Flag	✓	(✓)	Order
5	RogueN on Tokens [2]		✓	-	Overlap

Table 2.1: Comparator Overview [1]Lowercasing and removal of punctuation and whitespaces - latter not removed for token-based similarity metrics, [2]Removal of frequent tokens, * Preprocessing done by pre-implemented similarity measure, ✓ used in comparator, (✓) optional

2.2.2 Major Challenge - Named Entity Matching

As a first challenge in our project we had to find out that company names are inherently difficult to match. We found ourselves facing similar challenges as the Dutch Central Bank [1]. As with named entities in general every data source has a different level of detail, different data quality, and use of abbreviations among others. Regarding this, our matching rules had to cater to the following challenges:

1. **Company Name:** In part the name of the legal entity was used, in other cases the name of the group, and in other cases some abbreviation (e.g., Anheuser-Busch InBev Germany Holding GmbH vs. Anheuser-Busch InBev vs. AB InBev) or tokens of the name were omitted (e.g., Royal Dutch Shell vs. Shell). While all these names refer to the same entity, our matching rules evolved to address this.

2. **Data Quality:** We had to cope with general data quality issues which are represented for example by typos.

3. **Frequent tokens:** There are several tokens that have a higher frequency in company names. These include for example legal entity descriptors (limited, incorporated, ...), industy descriptors (bank, motors, pharmaceuticals, ...) and stop words (the, and, of ...). These may let names seem more similar than they actually are (General Motors vs. Hyundai Motors). We analyzed frequent tokens across our datasets and provided the matching rules with the option to remove frequent tokens.

4. **Token order:** Company names of different companies might be composed of similar tokens in a different order. Token-based similarity metrics alone would classify such names as similar altough they are not (Commercial National Financial vs. National Financial Group). This means that token order matters.

We adressed these challenges by combining comparators with different strengths in matching rules which we systematically evaluated at different final mathching thresholds. We also implemented some new similarity metrics to cater to our needs (Comparator 4 and 5). We also faced a second challenge concerning dataset sizes which will be outlined in 2.3.

2.2.3 Evaluation

We evaluated local and global matching strategies. In table XXX the best matching rule for each threshold is presented.

[1] https://medium.com/dnb-data-science-hub/company-name-matching-6a6330710334

2.3 Blockers

We used three different types of blockers (no blocker, symmetric, sorted neighborhood) and different blocking key generators. The following key generators were implemented, all with certain preprocessing options:

1. First letter of the company name.
2. Qgrams and first letter of the company name. [1] suggest this blocking technique to ensure a low degree of missed pairs while staying computationally efficient.
3. N starting characters of each company name token.

Especially the last blocker was born out of our second challenge, the large dataset size of the kaggle dataset.

2.4 Analysis of Errors

Chapter 3

Phase III - Data Fusion

	Baselines			Decision Tree			
Ontology	M(edian)	G(ood)	E(vil)	results	Δ-M	Δ-G	Δ-E
#301	0.825	0.877	0.877	0.855	+0.030	-0.022	-0.022
#302	0.709	0.753	0.753	0.753	+0.044	+0.000	+0.000
#303	0.804	0.860	0.891	0.816	+0.012	-0.044	-0.075
#304	0.940	0.961	0.961	0.967	+0.027	+0.006	+0.006
Average	**0.820**	**0.863**	**0.871**	**0.848**	**+0.028**	**-0.015**	**-0.023**

Table 3.1: Comparison between the Good and the Evil

If you cite something, do it in the following way.

- Conference Proceedings: This problem is typically addressed by approaches for selecting the optimal matcher based on the nature of the matching task and the known characteristics of the different matching systems. Such an approach is described in [**?**].
- Journal Article: S-Match, described in [**?**], employs sound and complete reasoning procedures. Nevertheless, the underlying semantic is restricted to propositional logic due to the fact that ontologies are interpreted as tree-like structures.
- Book: According to Euzenat and Shvaiko [**?**], we define a correspondence as follows.

These are some randomly chosen examples from other works. Take a look at the end of this thesis so see how the bibliography is included.

Bibliography

[1] Luis Gravano, Panagiotis G Ipeirotis, H V Jagadish, Nick Koudas, and T Labs. Approximate String Joins in a Database (Almost) for Free. page 10.

[2] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Appendix A

Program Code / Resources

The source code, a documentation, some usage examples, and additional test results are available at ...

They as well as a PDF version of this thesis is also contained on the CD

Appendix B

Phase II - Identity Resolution

-ROM attached to this thesis.

Appendix C

Further Experimental Results

In the following further experimental results are ...

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Master-/Bachelorarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Er- klärung rechtliche Folgen haben wird.

Mannheim, den 31.08.2014 Unterschrift