Multilingual Name Matching
Application to KYC


Marius Frunza
marius.frunza@schwarzthal.com


Ilgiz Mustafin
ilgiz.mustafin@schwarzthal.com


George Emelyanov
g.emelyanov@schwarzthal.com

Synopsis

Assessing the fully-fledged picture of a client is a crucial aspect of the KYC process. This assessment becomes complex when the client is resident and owns businesses in two or more countries. The compliance officer needs to find all the records for such clients in different databases and registries. When dealing with clients having data recorded in different languages and with different alphabets, the simple name matching methods have severe limitations.

For instance, a name like *Serge Pougacheff* can appear with very different forms when transliterated in Russian, English and French. Besides the original Cyrillic version Сергей Викторович Пугачёв, the Latin transcription includes: Sergei Pugachev, Serguei Pougacheff, Serge Pugachoff, Sergey Pugachyov, Serguei Pougatchev or Sergey Pugachew. Matching all possible forms of a name is the task of multilingual name matching, that is a driver in an efficient KYC process.

This paper analyzes a few optimal methods of multilingual entity matching used for entity resolution. The primary name matching approach using string comparison metrics is enriched with phonetic rules and with relational information. When applied in practice to solve KYC issues, most entity matching methods generate a significant amount of false positives. We enrich the current methods with a Bayesian approach based on the distribution of the frequency of name occurrence in a given language. The results applied to names of companies' directors from British and Russian business registries show that the approach using transliteration enhanced by phonetic matching and Bayesian search provides with the best performance

1. Introduction

Over the past five years, the banking sector has been hit by a wave of penalties for serious deficiencies in anti-money laundering and anti-terrorist financing systems and processes. These shortcomings are not solely the result of the propensity of some banks for customers or risky schemes. They also reflect the inefficiency of the compliance tools currently used by financial institutions, including KYC (Know Your Client). KYC is the process of a business identifying and verifying the identity of its clients.

Several banks (Table 1), including Danske Bank and Deutsche Bank, have suffered serious consequences, among other things because of shortcomings in their KYC methodology. Nevertheless, in many cases, the clients of these banks did not have a risky profile, but they had holdings in foreign companies, or they were associated with people subject to sanctions or involved in illegal activities. Therefore it is crucial in the KYC process to have a fully fledged picture of an individual, or a company with their corresponding transnational networks. The information concerning companies and directors are in many cases available on the national business registries. An advanced KYC should be able to access data from those sources and to map information about a person or entities from few different sources, containing data in different languages or different alphabets.

This process is called entity resolution also known as record linkage (Winkler (1999)), reference reconciliation (Saıs et al. (2007)) or object matching. It denotes the task of finding records from one or multiple databases, referring to the same real world entity Singla and Domingos (2006). Entity resolution in a single database case is sometimes called duplicate detection or deduplication Christen (2012).

As the amount of available data from national business registries increases entity resolution requires more resources and attention. In practice compliance manager in banks distinguish entities (person or companies) by themselves based on a manual process requiring human intervention. But for an efficient KYC process it is crucial to link entities across all business registries and data sources in order to provide compliance managers with advanced intelligence.

Name matching techniques are essential for joining data from different sources, especially from different business registries. Using exact string matching is not enough because one person can have multiple version of the names when transalted in different languages. As an example the name of the Russian ex-oligarch Sergei Viktorovich Pugachev appears in the Russian, English and French business registries in few versions. Beside the original Cyrillic version Сергей Викторович Пугачёв, the latin trascription include: Sergei Pugachev, Ser-

Bank/Country	Year	Penalty	Facts
Pilatus Bank (Malta)	2017	Liquidation	The bank has organized schemes to evade US sanctions against Iran.
HSBC (Hong Kong)	2018	0.35 Bil. USD	The bank has helped wealthy clients to avoid paying taxes..
HSBC (Hong Kong)	2012	1.9 Bil. USD	Cartels of drug traffickers laundered funds through bank branches in Latin America.
ABLV Bank (Latvia)	2018	Liquidation	The bank facilitated money laundering through illicit transactions for sanctioned entities in North Korea, Azerbaijan, Russia and Ukraine..
Deutsche Bank (Germany)	2010-2019	0.5 Bil. USD	The bank helped clients create offshore companies in tax havens to launder money.
Danske Bank (Danemark)	2007-2015	To be determined	The Estonian branch had transferred 235 billion USD largely to suspect customers in Russia and other former Soviet republics.
UBS (Switzerland)	2011-2013	0.02 Bil. USD	The bank's supervisory analysts would have cleared the alerts even as the transactions emitted warning signals.
US Bancorp (US)	2018	0.6 Bil. USD	The bank was investigating only a very limited number of suspicious transactions.
ING (Holland)	2010-2016	0.9 Bil. USD	The bank was convicted of non-prevention of money laundering and bribery, including bribery paid to the daughter of the Uzbek president by a unit of a Russian mobile phone operator.
BNP Paribas (France)	2004-2012	8.9 Bil. USD	The bank has put in place schemes that allow clients to circumvent sanctions against entities in Iran, Sudan and Cuba.

Table 1: List of the principal banks fines for KYC/AML inefficiencies.

guei Pougacheff, Serge Pugachoff, Sergey Pugachyov, Serguei Pougatchev. His son Alexander Pugachev appears also under Alexander Pugachew. Matching these possible forms of the name is the task of name matching, that is required in an efficient KYC process.

Applying name matching techniques has some difficulties. Usually, name similarity functions are designed to measure similarity between two words (two first names, two last names, etc.) and not between full names, thus names should be separated into parts and the corresponding parts should be found. Splitting full names into parts without knowing the context (language, naming customs of the person, etc.) without building a dictionary of all names is a hard problem.

Using name similarities is insufficient to assess whether two physical persons having the same name represent the same person or not. Other attributes can be used to solve this problem and to increase the accuracy of the matching overall. Moreover, the names that have a high frequency in a population like *Jean Dupont* in French, *Alex Jones* in English,*Weiten Li* in Chinese, *Alexander Ivanov* in Russian or *Aarav Patel* in Hindi generate massive amount of false positive in the name matching process.

In this paper, we explore the current approaches of name matching and examine their

validity. Then we discuss limitations of such approaches and introduce several techniques which can be used to overcome the limitations in the multilingual name matching. Later we show how an existing model can be improved with the new proposed techniques to do the multilingual entity resolution and present the performance.

This paper enriches the literature related to name matching methods applied to KYC. We discuss limitations of such approaches and introduce several techniques which can be used to overcome the limitations in the multilingual name matching. Later we show how an existing model can be improved with Bayesian search theory(Eisenstein et al. (2011), Sadinle (2017)) to reduce the false positives in the multilingual entity resolution and present the performance. The paper is organized as follows: *Section 2* discusses reviews the different strategies used for Entity resolution, *Section 3* focuses on the methods for assessing names similarities in different language, *Section 4* describes the resolution model framework introducing the Bayesian approach for matching estimation , *Section 5* presents the results of name matching methods applied for data from the British and Russian business registries. *Section 6* concludes.

2. Entity resolution methods

Entity resolution methods can be categorized by the level of using relational information in matching, as described in Bhattacharya and Getoor (2007):

- Attribute-only Entity Resolution. Record similarity depends only on the similarities of the attributes.
- Naive Relational Entity Resolution. Record similarity depends on the similarities of the attributes of the two entities (as in the previous case) and the similarities of the attributes of the records related to the two records being matched.
- Collective Relational Entity Resolution. Record similarity depends on the similarities of the attributes of the two records (as in the attribute-only case) and the similarity of the records related to the two records being matched.

All of the categories from above require an approach to estimate the similarity between the attributes of two entities can use different. For datasets concerning companies and persons the main attribute of an entity is the name. Thus, name maching is the key aspect of the entity resolution applied to KYC process. Given a name represented by the string A in one language and a name represented by the string B in possibly other language, a name

matching algorithm should tell if A and B represent the same person or give a probability of this (Peng et al. (2018); Patman and Thompson (2003)).

One alternative to exact string matching is to convert name strings to some common phonetic representation of the names and then to compare the phonetic representations. Some of the possible phonetic representations are Soundex Russell (1918), Match Rating Approach Moore (1977), Daitch-Mokotoff Soundex Mokotoff (2007), Beider-Morse Phonetic Matching Beider (2008), Double Metaphone Philips (2000).

- Soundex algorithm Russell (1918) is the ancestor of phonetic name matching algorithms. Soundex maps names to a special code consisting of a letter and three digits. The letter is the first letter of the name and the digits describes approximately the consonants of the name. The initial aim of Soundex was to be easily computed manually by human and was designed to be applied to paper documents. Its performance is relatively poor compared to more recent development.

- Match Rating Approach (MRA) Moore (1977) employs a very basic but straightforward process that transforms the string by deleting the vowels if the names does not start with a vowel and by deleting the second consonants in the pairs of double consonants, thereby reducing the name to a maximum of six characters by retaining the first and the last three characters. After the two names are encoded the MRA gives the matching rating of the two names based on the similarity of the two strings.

- Daitch-Mokotoff Soundex Mokotoff (2007) is an alternation of the original Soundex for Yiddish and Slavic languages. In this algorithm names are given six digit codes. The first letter is coded too (contrast to original Soundex where the first letter was retained as it was). Names can have multiple codes which is different from the original Soundex where names are mapped to only one code.

- Beider-Morse Phonetic Matching (BMPM) Beider (2008) was designed to decrease the number of false hits produced by Soundex-like algorithms. BMPM incorporates more than 300 common rules and has a number of language-specific rules to support 10 languages. The first step of BMPM is to identify the language and only then the approximate phonetic value is calculated based on the language detected.

After the names were converted to their phonetic representations (or if no conversion was done and the names remained as they were given), the two strings should be compared with a string distance metric like Levenshtein distance, Guth algorithm, Jaro Cohen et al. (2003), Jaro-Winkler Winkler (2006), etc.

- Levenshtein distance Levenshtein (1965) between two strings is a metric representing the number of one character changes (substitute a character by another character, remove a character, insert a character) needed to change one word into another.
- Guth algorithm was specifically designed to compare names. It takes two names as an input and as an output provides with the probability of the two names are variants of spelling of one name. Guth's algorithm compares two strings character by character, sometimes skipping or backtracking one or two characters.

3. Computing Name Similarities between English and Russian

In this section, we specifically discuss multilingual entity resolution problem with Russian and English names. It should be noted that we specifically chose Russian and English languages because of several reasons. The main reason is that the Russian language is highly phonetic (spelling represents pronunciation) and the English language is not so phonetic (spelling represents pronunciation poorly) and that the names from other languages (especially with Latin alphabet) are used in English in their original form (Sigmund Freud) rather than being transformed to better represent the original pronunciation (possibly Zikmont Froyd). As a consequence of this, we are actually having not only English-Russian name pairs, but also many other name pairs.

3.1. Transliteration

Russian and English languages use different alphabets (Cyrillic and Latin respectively) which makes string similarity functions (like Jaro-Winkler, Levenshtein) unsuitable for comparing names from these languages. To overcome this issue, we convert Russian names to Latin script.

In general, for a name there can be several possible valid equivalents in other language. Such conversion can be done via *transliteration*, *transcription* or *translation*. Transliteration is a more systematic and reversible procedure (Ельцин to Elcin), transcription is a more phonetic conversion focused on preserving the pronunciation of the name (Ельцин to Yeltsin). translation is done via mapping a name from the first language to a traditional equivalent name of the second language (Наталья to Nathaly), if there is one. All of these conversions have different rules and customs for different language pairs Li (2007), see 1.

Different domains can contain names produced by different conversion methods: Russian international passports have names converted from Russian to Latin script using a strict set of transliteration rules which are sometimes different from the transcription rules used in less formal contexts.

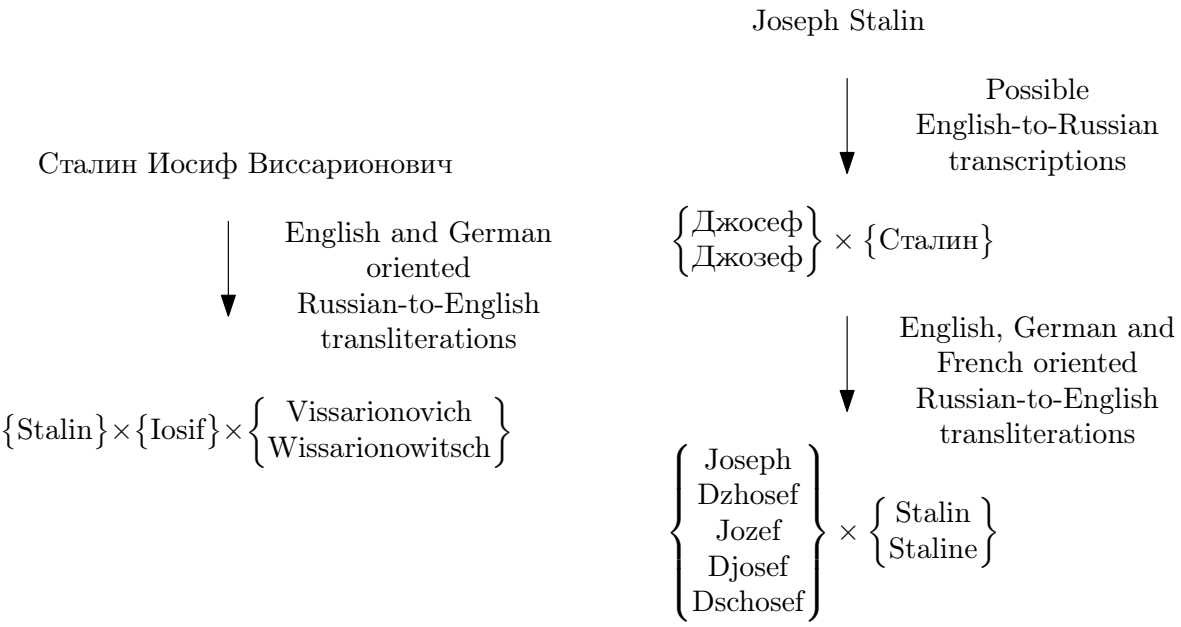

Figure 1: Cyrillic-Latin and Latin-Cyrillic name variations

Even though transcription can be used more widely, we will use the current Russian transliteration rules for the international passports because they are well defined.

3.2. Aligning Names

Names in different languages and cultures can have different structures: Russian names (e.g. Vladimir Vladimirovich Putin) have a given name (Vladimir), patronymic (Vladimirovich) and a family name (Putin), Spanish names (e.g. Pedro Sánchez Pérez-Castejón) have a given name (Pedro) and two family names(Sánchez and Pérez-Castejón), English names (e.g. Theresa Mary May) consist of the first name (Theresa), the last name (May) and optional middle names (Mary). Name variation can be entailed also by deviations of the same name in different languages.

Moreover, a name can appear in different forms: name parts can be omitted (Vladimir Vladimirovich Putin, Vladimir Putin), name parts can be reduced to initials (V. Putin), grammatical transformations like inflection (Vladimira Putina) and so on.

This variety of forms makes extracting name parts and computing name similarities a hard task. To partially overcome this issue, we introduce the *alignment* procedure as a part of computing name similarity.

Full-name similarity of two names with alignment is defined as the maximum similarity of all part permutations (V. Putin, Putin V.) with the second name. This way, similarity between Vladimir Putin and Putina Lyudmila will be high, but not maximum. It is not needed to

compute permutations of the second name, assuming the similarity function is symmetric.

3.3. Phonetic Transformations

Converting names between languages can make names less recognizable, especially so when translations or historical transliteration customs are used. For example, historically, letter "H" (sound $[\mathrm{h}]$) is transcribed from German to Russian as "Г" (sound $[\mathrm{g}]$). Consequently, converting Hermann to Russian and then to English might result in German. This is not the only example of such inconsistencies. Another example is the Persian name Rostam which is known as Rüstem in Turkish, Röstäm in Tatar, and Рустем in Kazakh languages. Converting to English by dropping diacritics Li (2007) will result in several different names: Rostam, Rustem, Rustam.

Sometimes names have equivalents for names in other languages. For example, Russian name Михаил (Mikhail) has a Ukrainian counterpart Михайло (Mykhailo). Both of these names would be printed in the passport in Soviet Union and the official transliteration would be Mikhail, but nowadays the name is spelled as Mykhailo Yermolovich (2001).

In the last section of the paper, we use Double Metaphone Philips (2000) to encode names. Double Metaphone was designed to account for the differences in writings of names in different languages. For example, Double Metaphone codes of both Michael and Maykl (from Майкл) will be equal to MKL.

4. Entity Resolution Model Description

Following the formalism introduce by Kouki et al. Kouki et al. (2017) we introduce the Probabilistic Soft Logic (PSL) Bach et al. (2015) framework. PSL uses *soft truth values* $\in [0, 1]$ and *relaxation rules* to encode logical models. Under the relaxation rules are Kimmig et al. (2012): PSL derives the objective function by translating logical rules specifying dependencies between variables and evidence into hinge-loss functions. PSL achieves this translation by using the Lukasiewicz norm and co-norm to provide a relaxation of Boolean logical connectives.

$$\begin{aligned} p \wedge q &= \max(0, p + q - 1) \\ p \vee q &= \min(1, p + q) \\ \neg p &= 1 - p \end{aligned}$$

As a result of training, the model will be able to the give probability of two mentions a and b referring to the same real-world entity: $\mathrm{Same}(a, b)$.

The proposed model consists of a number of PSL rules. The used rules can be separated into tow categories: attribute similarity rules and relational-attribute rules.

Attribute similarity rules state that if some attribute is similar in two references, then the references should be matched, and if some attribute is not similar in two references, then the references should not be matched. For example for two persons (director of companies a and b) we define:

$$\begin{aligned}
\mathrm{SimName}_{JW}(a,b) &\implies \mathrm{Same}(a,b)\\
\neg\,\mathrm{SimName}_{JW}(a,b) &\implies \neg\,\mathrm{Same}(a,b)\\
\mathrm{SimDOB}(a,b) &\implies \mathrm{Same}(a,b)\\
\neg\,\mathrm{SimDOB}(a,b) &\implies \neg\,\mathrm{Same}(a,b)\\
\mathrm{SameGender}(a,b) &\not\implies \mathrm{Same}(a,b)\\
\neg\,\mathrm{SameGender}(a,b) &\implies \neg\,\mathrm{Same}(a,b)
\end{aligned}$$

Where $\mathrm{SimName}_{JW}$ is the Jaro-Winkler name similarity of the two records, SimDOB is defined as maximum of the dates of births divided by the minimum of the two dates of births. $\mathrm{SameGender}(a,b)$ is a binary observed atom that takes its value from the logical comparison $a.gender = b.gender$ $\mathrm{Same}(a,b)$ is a continuous value to be inferred, which encodes the probability that the mentions a and b are the same person. To illustrate PSL in an entity resolution context, the following rule encodes that mentions with similar names, similar age and the same gender might be the same person:

$$\mathrm{SimName}(a,b) \wedge \mathrm{SimDOB}(a,b)\,\mathrm{SameGender}(a,b) \implies \mathrm{Same}(a,b)$$

Relational-attribute rules can be used together with the attribute similarity rules. Relational-attribute rules state that if two references mention similar companies (e.g. both have directors with similar names, both have the same address, both have similar sector of activity), then the two references should be matched. Given two companies x and y some relational-attributes rules can be:

$$\begin{aligned}
\mathrm{HasDirector}(x) \wedge \mathrm{HasDirector}(y) \wedge \mathrm{SimDirector}(x,y) &\implies \mathrm{Same}(x,y)\\
\mathrm{HasAddress}(y) \wedge \mathrm{HasAddress}(y) \wedge \mathrm{SimAddress}(x,y) &\implies \mathrm{Same}(x,y)
\end{aligned}$$

where $\mathrm{SimDirector}$ and $\mathrm{SimAddress}$ are defined as the maximum of Levenshtein and Jaro-Winkler similarities of directors' names and addresses respectively .

Let's assume for example that we would like to assess the linkage between a set of companies (A) from a business registry (ie United Kingdom) and another set of companies(B)

from another business registry (ie. Russia) For a relationship type t (*is a Director of* for $\mathrm{SimDirector}$) and two mentions α, β we find two sets $A = \{x : t(\alpha, x)\}$ and $B = \{y : t(\beta, y)\}$ which are the sets of mentions related to α and β respectively with the relationship t (for $\mathrm{SimDirector}$ A and B will be represented by the sets of all Directors α and β respectively).

Thus, the similarity between sets A and B will be the relational-attribute similarity of α and β and is defined as:

$$\mathrm{Sim}(A, B) = \frac{1}{|A|} \sum_{x \in A} \max_{y \in B} \mathrm{SimName}(x, y)$$

We can see that the name similarities estimates are used in many rules of the original model, thus the performance of matching depends heavily on the name similarity metrics.

However, the original definition of the $\mathrm{SimName}_J$ W metric has some limitations for applying it as-is to other datasets.

First, the underlying metrics compute similarities between first names, middle names, last names. Both of these functions are left undefined for the cases where names are represented as one string and it is unknown how to separate names into parts (first name, last name, etc.).

Second, Jaro-Winkler and Levenshtein distances which will give maximum distances for strings written in two different alphabets (e.g. Latin and Cyrillic). This also limits the applicability of the approach in multi-lingual context, and increase the dependency on the transliteration model used.

Third, for names having an increase frequency in a population, the similarity function employed in the framework above will generate a high number of false positive. Therefore, in order to reduce this bias it is necessary to take into account the infrequence of the occurrence of a name in the way the similarity functions is built.

When searching for the correspondence between names a and b the traditional metric can be alter in order to take into account the a priori frequency of the occurrence of the name b in the total population P, with a distribution denoted F_b, we introduce new similarity function:

$$\mathrm{SimName}^*_{JW}(a, b) = \sum_{b_i \in P} \mathrm{SimName}_{JW}(a, b) \cdot f(b, F_b) \tag{1}$$

depending on a Bayesian correction factor $f(b, F_b)$ which is penalizing the name with an increase frequency in the population.

5. Application to business registries data from Russian and United Kingdom

5.1. Dataset presentation

In order to assess the efficiency of the various entity resolution strategies and the accuracy of the name matching methods we built a relational database using data from the British (https://beta.companieshouse.gov.uk/) and Russian (https://egrul.nalog.ru/index.html) business registries . The databases contain information about companies, information about the companies' key persons and relationships indicated the role of the person in the company (ie director, manager)

The descriptive presentation of the dataset is exhibited in Table 5.1. The information about Russian companies and their key persons is presented in Russian language and written Cyrillic alphabet, while information about British companies and their key persons is in English written in Latin alphabet

Register	No of unique persons	No of unique companies	No of relationships	Attributes directors	Attributes companies
United Kingdom	2,388,638	1,000,262	4,425,058	Name, DOB, Address, Nationality	Name, Address, Tax number, date of incorporation
Russian Federation	293,655	169,180	324,210	Name, Tax number	Name, Address, Tax number, date of incorporation

Table 2: Presentation of datasets used in testing entity resolution algorithms

Given the set of key person in Russian companies we try to identify which are also key persons in British companies. This type of analysis would be useful in a KYC process when dealing for example with clients that could be involved in multi-national schemes of breaching sanction. The search is employing the methods described above and has the following steps :

- Data cleaning (including elimination of incomplete names or abnormal characters)
- Basic text processing (exclusion of punctuation or titles (Mr. Dr. Mrs. ....)
- Transliteration of the persons and companies' names from Cyrillic into English
- Name matching using various similarity function metric
- Entity resolution based on various strategies

5.2. Entity resolutions methods applied

Matching performances are measured with a set of different strategies for entity resolution and a set of algorithms for name matching (Mustafin et al. (2019)). For the entity resolution the strategies using PSL logic are considered:

1. *Name*, consiting in simple name matching
2. *Names + Personal Info (PI)*, consisting in name matching and matching of personal information (Address,Gender, Date of Birth)
3. *Names + Personal Info (PI) + Relational Info (RI) (1st degree: Company)*, consisting in name matching, matching of personal information and matching or related company names

The equivalence of full names (e.g. first name and last name in one string) which is the key step of all entity resolution strategies are assessed with name similarity functions. Various methods are employed :

In *Translit* matching, name similarities were computed as similarities between English and Russian versions of the names as described in 3.1. *Translit* matching was refined by by the alignment procedure described in 3.2 and the results are reported as *Translit Align*. *Translit Align* was further enhanced by phonetic matching techniques described in 3.3 and the results are reported as *Translit Align Phonetic*.

Two similarity function were used: the classic Jaro-Winkler metric and the Jaro-Winkler metric with a Bayesian correction.

Table 5.2 and 5.2 show the frequency of last and first names in the sample of the key persons for British and Russian companies, respectively. Form this analysis it appears that there is increase likelihood of matching correctly names like *David Smith* or *John Jones* but wrongly resolve the entity due to high occurrence of this first name / last name combination. In the case of our search looking for key person from Russian companies in the British registry it could be that *Aleksander Ivanov* or *Sergey Kuznetzov* may generate false positives. For this reason the Bayesian correction is necessary in order to deal with this cases

5.3. Results

The following metrics are used in order to assess the outcome of the resolution algorithms :

- *Precision* is the ratio of the true matches found to all found matches (i.e. what part of the found matches are true matches).

Last names			First name		
Name	Number	Frequency (%)	Name	Number	Frequency(%)
SMITH	35317	0.80	DAVID	130799	2.96
JONES	27319	0.62	JOHN	111393	2.52
BROWN	19580	0.44	MICHAEL	88622	2.00
WILLIAMS	19560	0.44	PAUL	78235	1.77
TAYLOR	17769	0.40	ANDREW	76391	1.73
DAVIES	15631	0.35	PETER	70445	1.59
PATEL	14225	0.32	ROBERT	63055	1.42
WILSON	12934	0.29	RICHARD	62991	1.42
THOMAS	12459	0.28	JAMES	62316	1.41
EVANS	12234	0.28	MARK	60208	1.36
KHAN	11920	0.27	STEPHEN	58922	1.33
JOHNSON	11025	0.25	CHRISTOPHER	57216	1.29
SINGH	10080	0.23	IAN	41098	0.93
ROBERTS	9487	0.21	SIMON	38681	0.87
ROBINSON	9373	0.21	WILLIAM	34118	0.77
WALKER	9329	0.21	ANTHONY	34036	0.77
THOMPSON	9199	0.21	NICHOLAS	30127	0.68
WHITE	9175	0.21	DANIEL	29103	0.66
HALL	9062	0.20	JONATHAN	29092	0.66
HARRIS	9035	0.20	MARTIN	29076	0.66

Table 3: Frequency of last and first names in the sample of the key persons for British companies

- *Recall* is the ratio of the true matches found to all true matches existing between the tw sets(i.e. what part of all matches was found).
- *F1 score* is the harmonic mean of precision and recall.

0 is the worst possible precision, recall, F1 score and 1 is the best possible.

The performance of the model on different data with different name matching techniques is shown in 5.

5.4. Discussion

The results show that the approach using transliteration enhanced by phonetic matching techniques provides globally with best results. The Bayesian correction absed on the distribution of name frequency improves massively the precision due to the decrease of false negatives. Several insights can be derived from these results.

Aligning and Phonetic transformation improves slightly the precision and recall. Aligning procedure described in 3.2 and 3.3) improves recall of matching by reordering words in

Last names			First name		
Name	Number	Frequency (%)	Name	Number	Frequency(%)
IVANOV	1544	0.48	ALEKSANDR	21151	6.52
KUZNETSOV	1054	0.33	SERGEY	19084	5.89
POPOV	870	0.27	VLADIMIR	13182	4.07
SMIRNOV	863	0.27	ANDREY	11435	3.53
IVANOVA	755	0.23	ALEKSEY	11321	3.49
PETROV	743	0.23	DMITRIY	9123	2.81
VASILYEV	709	0.22	YELENA	8222	2.54
KUZNETSOVA	522	0.16	NIKOLAY	7098	2.19
NOVIKOV	515	0.16	TATYANA	7064	2.18
MIKHAYLOV	513	0.16	YEVGENIY	7020	2.17
PAVLOV	506	0.16	YURIY	6536	2.02
SOKOLOV	502	0.15	IGOR	6417	1.98
MOROZOV	489	0.15	NATALYA	6228	1.92
KOZLOV	482	0.15	MIKHAIL	6123	1.89
VOLKOV	480	0.15	OLGA	6038	1.86
STEPANOV	479	0.15	VIKTOR	5778	1.78
MAKAROV	457	0.14	OLEG	5549	1.71
FEDOROV	449	0.14	IRINA	5278	1.63
SEMENOV	449	0.14	SVETLANA	4792	1.48
YEGOROV	447	0.14	VALERIY	4178	1.29

Table 4: Frequency of last and first names in the sample of the key persons for Russian companies

names to increase the similarity. If the first name is mentioned just by the initial (ie O for Oleg) this is also taken into account compared to the basic name matching. However, aligning can increase the similarity of actually non-matching names thus decreasing the precision of matching.

This improvement encompasses variations of the transliteration mentioned in the sections above. For example *Oleg Morozov* is coreclty matched against the Latvian version *Olegs Morozovs*.

Using Double Metaphone codes for computing similarities (described in improved the precision of matching by putting several different transliterations of one name into same

Feature Set	Precision	Recall	F1 Score
Translit Names	.1	.58	.17
Translit Names Align	.15	.59	.24
Translit Names Align Phonetic	.16	.59	.25
Translit Names + PI	.11	.6	.19
Translit Names Align + PI	.17	.63	.27
Translit Names Align Phonetic + PI	.63	.71	.27
Translit Names + PI + R1	.79	.15	.25
Translit Names Align + PI + R1	.79	.15	.25
Translit Names Align Phonetic + PI + R1	.79	.15	.25

Table 5: Performance for various strategies of entity resolution and various name matching techniques based on classic similarities metric

Feature Set	Precision	Recall	F1 Score
Translit Names	.6	.59	.059
Translit Names Align	.61	.61	.61
Translit Names Align Phonetic	.61	.61	.61
Translit Names + PI	.62	.62	.62
Translit Names Align + PI	.63	.63	.63
Translit Names Align Phonetic + PI	.63	.65	.64
Translit Names + PI + R1	.79	.15	.25
Translit Names Align + PI + R1	.79	.15	.25
Translit Names Align Phonetic + PI + R1	.79	.15	.25

Table 6: Performance for various strategies of entity resolution and various name matching techniques based on similarities metric with Bayesian correction

bucket e.g. For example *Aleksander and Olexander* (a transliteration of Александр both have the same Double Metaphone which results in the maximum name similarity.

Accounting for personal information as an entity resolution attributes improves slightly the

precision and recall. Personal information is available only for British companies' key persons and less available for Russian data. But in many cases the gender of the person can be implied from the title (Mr. vs, Mrs) in UK and for the termination of the last name for Russian entities (..ov vs ..ova).

Accounting for relationship information as an entity resolution attributes improves massively the precision and but reduces the recall. Accounting for relationship information will look for people with similar names that are involved in companies with similar name (ie British Petrol LTD and OOO British Petrol). This reduces the likelihood of hitting false positives, but reduces the recall as the matches represent only a small part of all true matches.

False positives are generate for non-Russian, Non-English names. Persons with Chinese/Asian names like Yen Chan Sen or San Bo Li that have companies in Russia generate false positive when matching persons from the British registry.

The Bayesian correction of the similarity metric improves all metrics. While less frequent names as Oleg Deripaska are correctly matched by all algorithms, the frequent name are the root cause of false positive. This is also due to the fact that in United Kingdom the authorities do not register systematically the patronymic name, which would improve the match.

6. Conclusions

This paper explores methods of multilingual entity matching applied to KYC . The basic name matching approach using string comparison metrics is enriched with phonetics rules and with relational information. The results show how different techniques and different entity resolution strategies affect precision and recall. A Bayesian correction is applied to classical similarity metrics in order to account for the frequency of occurrence of a name. It provides with better precision. Nevertheless, the presented approaches do not fully address the issues of name variability in the transliteration/transcription/translation process between languages with different alphabets. Our results highlight the need for a change of paradigm by replacing the static and one-dimension concept of Know Your Client (KYC) with a dynamic, multi-dimensional and forward-looking concept of "Know Your Network" (KYN). This aspect will be explored in a further research.

References

Bach, S.H., Broecheler, M., Huang, B., Getoor, L., 2015. Hinge-loss markov random fields and probabilistic soft logic. arXiv preprint arXiv:1505.04406 .

Beider, A., 2008. Beider-morse phonetic matching: An alternative to soundex with fewer false hits. Avotaynu: the International Review of Jewish Genealogy .

Bhattacharya, I., Getoor, L., 2007. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD) 1, 5.

Christen, P., 2012. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media.

Cohen, W., Ravikumar, P., Fienberg, S., 2003. A comparison of string metrics for matching names and records, in: Kdd workshop on data cleaning and object consolidation, pp. 73–78.

Eisenstein, J., Yano, T., Cohen, W.W., Smith, N.A., Xing, E.P., 2011. Structured databases of named entities from bayesian nonparametrics, in: Proceedings of the First workshop on Unsupervised Learning in NLP, Association for Computational Linguistics. pp. 2–12.

Kimmig, A., Bach, S., Broecheler, M., Huang, B., Getoor, L., 2012. A short introduction to probabilistic soft logic, in: Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications, pp. 1–4.

Kouki, P., Pujara, J., Marcum, C., Koehly, L., Getoor, L., 2017. Collective entity resolution in familial networks, in: Data Mining (ICDM), 2017 IEEE International Conference on, IEEE. pp. 227–236.

Levenshtein, V., 1965. Binary codes capable of correcting spurious insertions and deletion of ones. Problems of information Transmission 1, 8–17.

Li, C.W.C., 2007. Foreign names into native tongues: How to transfer sound between languages—transliteration, phonological translation, nativization, and implications for translation theory. Target. International Journal of Translation Studies 19, 45–68.

Mokotoff, G., 2007. Soundexing and genealogy. URL: http://www.avotaynu.com/soundex.html .

Moore, G.B., 1977. Accessing individual records from personal data files using non-unique identifiers. volume 13. US Department of Commerce, National Bureau of Standards.

Mustafın, I., Frunza, M.C., Lee, J., 2019. Multilingual entity matching, in: International Conference on Advanced Information Networking and Applications, Springer. pp. 810–820.

Patman, F., Thompson, P., 2003. Names: A new frontier in text mining, in: International Conference on Intelligence and Security Informatics, Springer. pp. 27–38.

Peng, T., Li, L., Kennedy, J., 2018. A comparison of techniques for name matching. GSTF Journal on Computing (JoC) 2.

Philips, L., 2000. The double metaphone search algorithm. C/C++ users journal 18, 38–43.

Russell, R., 1918. Index. URL: https://www.google.com/patents/US1261167. uS Patent 1,261,167.

Sadinle, M., 2017. Bayesian estimation of bipartite matchings for record linkage. Journal of the American Statistical Association 112, 600–612.

Saıs, F., Pernelle, N., Rousset, M.C., 2007. L2r: A logical method for reference reconciliation, in: Proc. AAAI, pp. 329–334.

Singla, P., Domingos, P., 2006. Entity resolution with markov logic, in: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE. pp. 572–582.

Winkler, W.E., 1999. The state of record linkage and current research problems, in: Statistical Research Division, US Census Bureau, Citeseer.

Winkler, W.E., 2006. Overview of record linkage and current research directions, in: Bureau of the Census, Citeseer.

Yermolovich, D., 2001. Imena sobstvennyye na styke yazykov i kultur [proper names across languages and cultures]. Moscow: R. Valent .

SCHWARZTHAL TECH

Know You Network, AI meets KYC

Visit schwarzthal.tech for more

Marius Frunza
marius.frunza@schwarzthal.com

Ilgiz Mustafin
ilgiz.mustafin@schwarzthal.com

George Emelyanov
g.emelyanov@schwarzthal.com

Contact

contact@schwarzthal.com
FR: +33627297834
UK: +447952208734
RU: +79655142975

Address

211 Business Design Center
London, N1 0QH,
United Kingdom

Social

Vk.com
/schwarzthal_tech
Twitter
@schwarzthal
Linkedin
/company/schwarzthal-tech