ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0081

Cross-Language Name Matching for Data Fusion in Linked Open Data

Ziad F. Torkey
Computer Science
Arab Academy for Science and Technology
Cairo, Egypt
ziadtorkey@gmail.com

Emad Elabd
Faculty of Computers and Information
Menoufia University
Menoufia, Egypt
emadqap@gmail.com

Mostafa Abdelazem
Faculty of Computer Science
Arab Academy for Science and Technology
Cairo, Egypt
melbaqary@gmail.com

Abstract -- Data quality and accuracy affects the success of data integration in Linked Open Data (LOD). The main goal of data fusion is to represent each real-world entity once on the Web. Data inaccuracy problems exist due to misspelling and a wide range of typographical differences mainly in non-Latin languages, those problems become more complicated when a person is identified by a name, and this name can be presented differently in same/different languages. Up to author's knowledge, the previous approaches which supported Arabic person names are not designed to work with LOD. This paper proposes a framework that uses person names as matching criteria from cross-language LOD Datasets. The proposed framework has substantial improvements in matching results compared to state of the art framework of matching techniques with better matching rate which exceed 6% in precision and 6% in recall.

Key words—Data fusion; ontology Alignment; duplicate detection; linked open data; semantic web.

I. INTRODUCTION

Information technology plays an important role in today's IT based economy. Many industries and systems depend on the accuracy of data to carry out operations [1]. In the typical Web (Web 2.0), there are links between documents and the relationship between any linked documents is implicit. Sometimes the information in the web is redundant and the same data has multiple representations [2].

Due to the redundancy of real-world entities over the web, the idea of URI (Unified Resource Identity) was presented in the Linked Open Data (LOD) [2]. Linked Open Data (Web 3.0) represents the same real-world object into unique identity and consistent representation [3].

LOD is a collection of Ontology [2] published over the Web to present things uniquely. Ontology are released in the form of resource description framework (RDF). Ontology over the LOD contains millions of RDF triples (subject, predicate and object). LOD elevated links between different datasets/data sources which characterized the relations between things to facilitate browsing for users [1].

Things in different LOD ontology are presented like (Companies, food, persons) as datasets. Dataset producers like Dbpedia [4] publish datasets for different categories of things based on the available data they have. This leads to the problem of presenting the same real-world entity more than once with different data available on each source of data [3].

doi:10.15849/icit.2015.0081

The quality of the data stored in LOD can have significant cost effect on a system that uses the information to conduct business. Data fusion is needed in LOD applications to enhance the quality of data [3]. The main goal of data fusion is to integrate different data which represent the same real-world objects and the resolution of data conflicts. The quality of data can be affected by many factors including spelling mistakes, errors in data entry and different conventions in storing information. For example, Arabic data has more problems than Latin based language (English, French and German) because of these different conventions [5].

Arabic is the main language for millions of people in twenty Middle East and North African countries [6] [7]. Arabic language has characteristics like absence of capital letters, complex morphology and short vowels [8]. Since Arabic is one of the languages used in the published Datasets, Data fusion is used for matching things presented in different data sources using different languages.

One category of the published datasets in LOD is datasets which present persons information. Datasets are produced by different vendors all over the Web. Person's datasets present all available data that can be found about the person like (Name, Age, Work in, Birthdate, etc.). Names can be used as matching criteria for those persons across different data sources. Since the same person can be presented in different data sources in different languages (English- Arabic), this means they can be matched using his/her name and any other available data for this person.

Names written in English cannot directly be matched to names written in Arabic due to different language script and morphology. We propose in this paper a matching framework that is based on phonetic techniques to match person names across different sources in different languages (English – Arabic) so that a single person is presented once on the LOD.

The rest of the paper is organized as following: Section II demonstrates the problem in name matching across different languages (English – Arabic). Section III presents an overview about the work done in Ontology alignment field. Section IV describes the proposed framework and how it can help in improving the matching results between English and Arabic names. Section V shows the impressive result using the proposed framework and also in comparison to latest frameworks available. Section VI concludes the paper and the future work.

II. PROBLEM DEFINITION

Data fusion is one of the biggest problems in providing a trusted source of data [2]. Data fusion is needed to achieve the main objective of Linked Open Data, which is presenting a real-world entity once with a single unified resource identity (URI).

One of the problems of data redundancy over the LOD is Person's data. Person data can be redundant due to misspelling or different presentation in different languages (English –Arabic) over the LOD [9]. Same person name cannot be matched in two different datasets written in different languages like (English dataset- Arabic dataset). In addition to that datasets can have misspelled person names in both languages which increase the difficulty of fusing person data.

Therefore, the contributions of this paper are significant for many reasons. Firstly it proposes an automated technique that enhances string matching between multiple data sources containing redundant data. Secondly the framework has the ability to matching person data in cross-language (English-Arabic) using person names as matching criteria. Finally the framework preforms person names matching on different ontology in LOD.

III. RELATED WORK

Data fusion is a problem which still needs a lot of work to be done [3] [10] [11]. Work has been accomplished to propose fundamental techniques for string matching [12] [13]. String matching is a classical problem which has been there before known in databases integration [1]. For more than five decades, the traditional database community has discussed this problem and a lot of work has been done [14].

String matching techniques can be grouped into three classes [13]: global versus local, set versus whole string and perfect-sequence versus imperfect-sequence [13], the first class refer to the amount of information the technique needs to classify a pair of strings as a match or no-match, global techniques start with computing information over the string labels in ontology triples before it matches any strings, in local techniques the string pairs are being considered as the only input required, examples of this class are:

- TF-IDF: this technique is based on that two strings are similar if they share a word that is rare in the ontology [15].

doi:10.15849/icit.2015.0081

- Soft TF-IDF: this technique is based on Jaro Winkler technique which works on words equality rather than exact match.

The second class consists of two sub-classes, perfect-sequence which requires characters in the pair of strings to occur in the same order so it can be considered as match, imperfect-sequence is using the same technique as the perfect-sequence with a relaxing condition based on a threshold, this condition increases the false match's rate, and examples of this class are:

- Jaccard: The number of words in pair of strings which are having common characters divided by the total number of the unique words in each string.
- RWSA (Redundant, Word-by-word, Symmetrical, and Approximate): strings characters are replaced by their Soundex code, there is a match if each word in the shorter string has a weighted edit distance less than a threshold from a word in the longer string [13].

The third class works by finding the overlap between pair of strings, it works better on long strings, examples of this class are:

- Levenstein [12]: the number of substitutions needed to transform one string to another.
- N-gram [16]: string is converted to a set of n-grams, the results are compared using similarity metric.

Many techniques are explored including machine learning. Some frameworks use training data to semi-automatically find an entity matching strategy to solve a match problem. The quality of the computer string matching process is found to be higher than the manually linked record (done by humans) [17]. TAILOR [18] is a flexible record matching toolbox which allows the users to apply different duplicate detection methods on the datasets. BigMatch [19] is a duplicate detection program which is used by the US Census Bureau. If the sizes of the datasets are large, online record linking can be used [10]. FEBRL [20] is one of the tools that perform record linkage/duplicate detection process. FEBRL includes a new approach for improved data cleaning and standardization that support parallelization [21]. FEBRL needs to be installed on a local machine and configure the operating system and prerequisite software to match FEBRL platform requirements, which is not suitable for use on the web.

DRDAA (Duplicate Record Detection with Arabic adjustment) is the latest Web-based matching framework that supports cross-language string matching [22]. DRDAA has predefined rules which have been set by subject expert matter in the field of Arabic especially Arabic names. DRDAA has the ability to find matching person names from two different data sources from different languages (Arabic-English). DRDAA is based on rules which have been based on human experiment, which means that the framework is limited to expert's knowledge.

Up to our knowledge and experiments with the current available frameworks and tools, most of them does not support the matching of Arabic names, and none of them support matching cross-language names in Linked Open Data.

IV. CROSS-LANGUAGE NAME MATCHING FRAMEWORK

The proposed framework is a Web-based string matching based on person names in cross-language. Cross-language Name Matching Framework is designed and implemented to overcome the missing feature of names matching in Linked Open Data. The architecture of the proposed framework consists of Datasets selection, Names Triples Listing, Data cleaning and standardizing, creating phonetic coding and finally Name matching and linking as shown in Figure 1.

Datasets from cross-languages (Arabic-English) in Linked Open Data are selected as an input for the framework. Triples that contain person's names are used for matching. Data cleansing and standardization is required for insuring the quality of data for matching. The proposed framework provides name matching between Datasets that have redundant data about persons. Datasets can be fused by matching the names of persons using Soundex and ASoundex techniques and creating new triples for Soundex code values. Soundex code triples gets compared for matching and when a match is found a new SameAs triple is created between the two datasets showing the equality between the two entities (persons) in both datasets.

doi:10.15849/icit.2015.0081

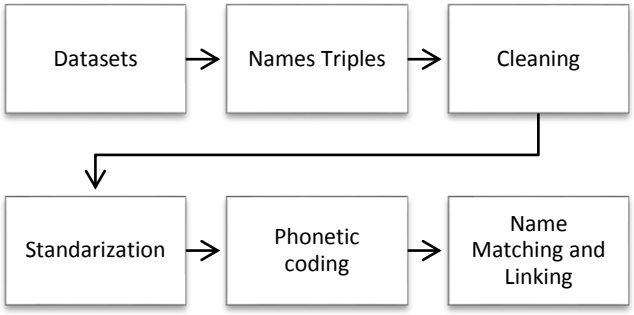

Fig.1. Cross-language Name Matching Architecture

The following sub-sections discuss the framework in details.

A. *Data sources*

Person's datasets were selected from FOAF:Person and yago **[23]** datasets. Two datasets were used as an input for the framework, first in English and second in Arabic. Person datasets contain data like (Name, Age, Birthdate, etc.). Datasets are converted into RDF Graphs **[2]** so person data is represented in a form of triples as shown in figure 2. Person name triples are selected for cleaning and standardizing.

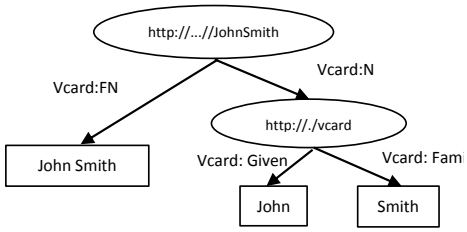

Fig.2. Example on RDF Triples

B. *Data Cleaning and Standardizing*

The framework uses triples containing person names as an input for cleaning process. Data cleaning is the process of removing all inconsistent or rubbish data like (null triple, "aaaaa", "___", "!!!!!"), example on inconsistent data is finding only numbers in name property or finding characters in an integer property like date or age.

In the datasets names might be stored with prefixes. An example for this problem is "prof. ahmed", these names can be represented in different ontology as "ahmed", "استاذ احمد", "احمد" and all of them represented the same real-world object. After studying number of data samples, sometimes names are written with a prefix like ("Dr. Mostafa", "Eng. Magdy", "junior", "أ. عماد", "السيد طارق"). Those prefixes were collected in following table:

Table 1: Names prefixes in English and Arabic

English Prefix	Arabic Prefix
Dr.	دكتور
Prof Dr	استاذ دكتور
Prof.	استاذ
Prof	أستاذ
Eng.	م.
Eng	مهندس

Prefixes are removed from names properties in the selected triples so the remaining string in the object is the name without any distraction. Finally the name string is trimmed to remove any unrequired spaces in the string.

One of the problems in person names in Latin based languages is the multiple representation of a word [12]. In Arabic language, the problem may occur in one character like "أ" which can be represented as "ا,أ,آ" and this character using will be based on the pronunciation. Data standardization is used to unify the Arabic Dataset so that misspelling or different pronunciation can be controlled and unified. Set of standardization rules shown in Table 2 are applied on Arabic dataset.

Table 2: Standardization rules

Set of characters	Equivalent character value
أ ، إ ، آ ، ا	ا
ي ، ى	ى
ه ، ة	ه
و ، ؤ	و

C. *Phonetic coding*

Previous approaches worked on aligning ontology in different languages base on translation [24]. Person names cannot be translated, if an Arabic name like "سعيد" (Seed) was translated into English it would be "happy" which is not the same meaning. The proposed framework converts names into phonetic code. This phonetic code should be

doi:10.15849/icit.2015.0081

equivalent in any language when phonetic technique is used.

Soundex was invented by Russell [25], it is the most common phonetic coding scheme for Latin-based languages, and it is based on replacing characters with phonetic code. ASoundex [26] was introduced later for Arabic language which followed the same pattern of Soundex with a little bit of tweaks.

The proposed framework uses Soundex and ASoundex for creating the phonetic code. String name value found in name, full name or given name property. The selected list of triples is converted into Soundex/ASoundex code and stored as new triple attached to the person URI, later on new Soundex or ASoundex property triples get attached back to the original Graph. This Soundex/ASoundex code property is used as matching criteria between datasets from two different languages (English-Arabic) which may contain redundant data about persons.

Based on the following table the initiation of the Soundex code is created:

Table3: Initiation of Soundex and ASoundex Code

Code	Characters	English phonetic equivalent	Category
1	ب،ف	b,f	Labial
2	خ،ج،ز،س،ص،ظ،ق،ك	k,q,z,s,c,z,j,kh	Guttural and sibilants
3	ت،ث،د،ذ،ض،ط	t,d	Dental
4	ل	l	Long liquid
5	م،ن	m,n	Nasal
6	ر	r	Short liquid

Problems were found in English names like “Charly”, it can be pronounced in Arabic as “شارلي” so that we can consider “Ch” as “ش” in Arabic, but when we ran to a name like “Christen” we found that “Ch” is considered as “ك” in Arabic. Another problem was in an Arabic name like “اسامة”, this name can be written in English as “Osama” or “Usama”, both pronunciations are correct. With using Soundex technique first character gets reserved which mean some times it will be “U” and other times it will be “O”. Some conditions needed to be added to the framework to solve those problems.

D. Name Matching

Soundex code triples from English and Arabic Graphs get compared looking for similarities. When equal code triples are found a new owl:SameAs triple gets created between the two entities from both Graphs (English – Arabic) as shown in figure 3.

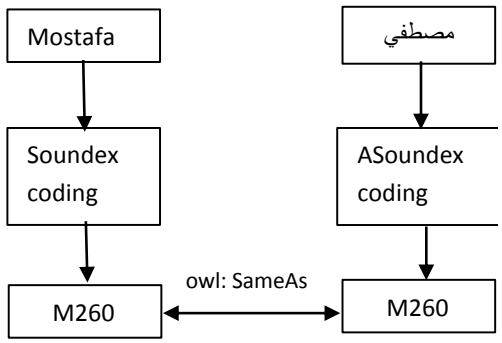

Fig.3. Example of Name Matching using Soundex and ASoundex coding

V. EXPERIMENT AND RESULTS

For testing the framework, English and Arabic datasets were selected to perform number of experiments to check the performance of the framework. Based on the authors’ knowledge, DRDAA framework was chosen for comparing the performance on the proposed framework.

A. Experiment 1: Sample of 100 names triples.

In this experiment we used English and Arabic datasets containing 100 person data. The table below shows the results of precision and recall for this experiment.

Table 4: Experiment 1 results.

Quality metric	Proposed Framework
No. of Entities	100
True Positives (TP)	79
True Negatives (TN)	0
False Positives (FP)	2
False Negative (FN)	19
Precision (TP/(TP+FP))	98%
Recall (TP/(TP+FN))	81%

Experiment 1 results 98% in precision and 81% in recall. After investigating these results, we found that some English names can be pronounced differently if it starts with a special sequence of characters. Examples for those characters are (“C”+”H”+”R) and (“C”+”H”+”A”). Some tweaks needed to be added to the standard Soundex technique to overcome this problem.

doi:10.15849/icit.2015.0081

B. *Experiment 2: Sample of 100 names triples*

In experiment 2, same sample of person data used in experiment 1 were used in experiment 2. The table below shows the results of precision and recall for this experiment after modifying the Soundex technique.

Table 5: Experiment 2 results

Quality metric	Proposed Framework
No. of Entities	100
True Positives (TP)	96
True Negatives (TN)	0
False Positives (FP)	1
False Negative (FN)	3
Precision (TP/(TP+FP))	98 %
Recall (TP/(TP+FN))	96 %

Experiment 2 gave significant results with 98% in precision and 96% in recall.

C. *Experiment 3: Comparison between DRDAA and the proposed cross-language name matching framework.*

In this comparison, 3 thousand triples of English and Arabic person names extracted from FOAF: Person [27] and yugo [23] datasets were used for testing the framework. Comparing the proposed framework with the latest similar framework for string matching which is DRDAA [22] we found an improvement in the results as shown in Table 6.

Table 6: Comparison between results between DRDAA and proposed framework

Quality metric	DRDAA	Proposed Framework
No. of Entities	3000	3000
True Positives (TP)	2415	2742
True Negatives (TN)	0	0
False Positives (FP)	210	39
False Negative (FN)	375	219
Precision (TP/(TP+FP))	92%	98%
Recall (TP/(TP+FN))	86%	92%

Comparing the proposed framework results with DRDAA framework which is the state of the art in data fusion and record linkage we found that we have the advantage of higher matching rate which exceed 6% in precision and 6% in recall. The proposed framework is fully automated which is an advantage over the DRDAA that is based on subject expert matter experience and that can be a limitation for this framework. Finally the proposed framework is the only person name matching approach that is available for data fusion in Linked Open Data.

VI. CONCLUSION AND FUTURE WORK

Data fusion is an important step in Ontology alignment. In this paper, a web-based framework for cross-language name matching in LOD is proposed with enhanced phonetic technique. The proposed framework helped in fusing data conflicts and redundancy over LOD Datasets. In the future we will work on new phonetic technique that takes in consideration the pronunciation and the punctuation of names which will increase the precision and recall rate and give much better results in cross language matching.

REFERENCES

[1] J. Zhu, "Duplicate Record Detection," in *ElSevier*, 2012.

[2] T. H. C. B. Tim Berners-Lee, "Linked Data - The Story So Far," 2009.

[3] F. N. Jens Bleiholder, "Data Fusion," in *ACM*, 2008.

[4] T. T. Jr, "dppedia," Wikipedia, 2015. [Online]. Available: http://dbpedia.org/.

[5] S. C. V. G. Rohit Ananthakrishna, "Eliminating Fuzzy Duplicates in Data," in *ACM*, 2002.

[6] Berners-Lee, "Semantic Web Road Map," in *W3 organization*, 1998.

[7] L. &. A.-K. Saleh, "AraTation: An Arabic Semantic Annotation Tool," 2009.

[8] A. R. A. R. I. Majdi Beseiso, "A Survey of Arabic Language Support in Semantic Web," in *International Journal of Computer Application*, 2010.

[9] M. H. P. Cheatham, " The role of string similarity metrics in ontology alignment," in *Tech. rep., Kno.e.sis Center*, 2013.

[10] D. V. M. a. L. D. Dey, "Efficient Techniques for Online Record Linkage," in *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 2011, pp. 373-387.

[11] P. H. A. P. S. K. V. a. P. Z. Y. Prateek Jain, "Ontology Alignment for Linked Open Data," in *Springer*, 2010.

[12] D. S. L. C. a. Dr. Andrew T. Freeman, "algorithm, Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance," in *ACM*, 2006.

[13] M. C. a. P. Hitzler, "String Similarity Metrics for Ontology Alignment," in *Kno.e.sis Center, Wright State University*, USA, 2014.

[14] P. G. I. V. S. V. Ahmed K. Elmagarmid, "Duplicate Record Detection," in *IEEE*, 2007.

doi:10.15849/icit.2015.0081

[15] R. W. P. L. K. F. W. Ho Chung Wu, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, June 2008, p. 13, 2008 .

[16] P. V. d. V. J. D. P. R. L. M. Peter F. Browen, "Class-Based n-gram Models of natural Language," in *ACM*, 1992.

[17] P. K. ,. C. G. a. J. N. Wilbert Heeringa, "Evaluation of string distance algorithms for dialectology," in *Linguistic Distances*, Sydney, Association for Computational Linguistics, 2006, pp. 51-62.

[18] M. V. V. a. A. E. Elfeky, "TAILOR: a record linkage toolbox," in *Proceedings of the 18th International Conference on in Data Engineering*, 202.

[19] W. Yancey, "Bigmatch: A Program for Extracting Probable Matches from a Large File for Record Linkage," Bureau of the Census, US, 2002.

[20] P. Christen, "Febrl: a freely available record linkage system with a graphical user interfac," *the second Australasian workshop on Health data and knowledge management* , vol. 80, no. Australian Computer Society, pp. 14-25, 2008.

[21] P. T. C. a. M. H. Christen, "Febrl – A Parallel Open Source Data Linkage System," in *Advances in Knowledge Discovery and Data Mining*, Berlin, Springer, 2004, pp. 638-647.

[22] A. H. Y. A. H. Tarek El Tobely, "Web-based Arabic/English Duplicate Record Detection with Nested Blocking Technique," in *IEEE*, Cairo, 2014.

[23] datahub, "Yago," CKAN , 2013. [Online]. Available: http://datahub.io/dataset/yago.

[24] S. e. a. Zaidi, "A Cross-language Information Retrieval: Based on an Arabic Ontology in the Legal Domain," in *International Journal of Computer Applications*, 2009.

[25] Russel, "Soundex". USA Patent 1261167, April 1918.

[26] S. B. E. J. D. G. a. O. F. Syed Uzair Aqeel, "On the Development of Name Search Techniques for Arabic," in *WILEY interScience*, Chicago, 2006.

[27] L. M. Dan Brickley, "FOAF Project," FOAF, 2014. [Online]. Available: http://www.foaf-project.org/.

[28] R. G. B. M. D Brickley, RDF Vocabulary Description Language 1.0: RDF Schema, W3C, 2004.

[29] F. v. H. Deborah L. McGuinness, owl web ontology language overview, W3C, 2004.