Tensor Factorization for Cross Lingual Entity Co-reference Resolution in the Linked Open Data

Melkamu Beyene
Addis Ababa University
Addis Ababa, Ethiopia
melkamu.beyene@aau.edu.et

Pierre-Edouard portier
INSA de Lyon
Lyon, France
Pierre-Edouard.portier@insa-lyon.fr

Solomon Atnafu
Addis Ababa University
Addis Ababa, Ethiopia
solomon.atnafu@aau.edu.et

Sylvie Calabretto
INSA de Lyon
Lyon, France
sylvie.calabretto@insa-lyon.fr

ABSTRACT

The main objective of this research was to identify co-referent entities located in several linked open data (LOD) sources that are described in various natural languages. The problem is approached from two perspectives. First, we do a multi-scale analysis of the RDF graph to discover structural similarities of entities. This was implemented as a tensor decomposition of the RDF graph with each predicate corresponding to a horizontal slide of the tensor. Hereafter, we used the term "structural evidence" to refer to the result of this analysis. Second, for each entity, we associated textual data coming from the Web of documents. Thus, after some preprocessing (viz. removing empty words, applying a weighting scheme such as tf-idf, ...), we represented each entity in a high dimensional space with each dimension corresponding to a term. Next, through a Singular Value Decomposition (SVD), we find a subspace such that the sum of squared distances from the original space to the sub space is minimized. This dimensionality reduction allows us to find language independent similarities between entities. Hereafter, we use the term "textual evidence" to refer to the result of this analysis.

Since the similarity information coming from the structural and the textual evidence are complementary to each other, a global similarity score is computed by aggregating the two evidences. We adopt a linear opinion pool, an approach which is commonly used in belief aggregation as an aggregation mechanism. In the end, for any given entity, we obtained a global similarity vector. The higher component values of this vector correspond to potential co-referent entities.

All algorithms are implemented in Python. According to the experiment result conducted on the French and English DBpedia, our approach can bring high results.

Keywords

Linked data; cross lingual entity linking; tensor factorization; Semantic web; entity co-reference resolution, linked open data

1. INTRODUCTION

Due to the linked open data (LOD) project, billions of RDF data sets are published. The main goal of the project is to provide extensive links between LOD sources at the level of instances [1]. Instance level links enable people or software agents to walk from one data source to others in order to obtain a relatively comprehensive view of the entities [2].

Interlinking resources that represent the same real-world object across linked open data sources is a widely researched topic [2, 3, 4, 5, 6]. However, all of them were evaluated on monolingual data sets. Recently, the LOD cloud has shown an increase in resources published at various natural languages. Although the importance of linking instances across language has been argued in several research works [32, 33], it has not been well studied. To the best of our knowledge, interlinking of multilingual LOD instances has been tested only by translating language specific descriptions of entities through machine translation [18,31]. However, this kind of approach is highly dependent on the quality of the translation system which is scarcely available in many natural languages [19].

There are challenges that need attention when one wants to identify co-referent entities in multilingual linked open data (MLOD). The first challenge is the increase of schema heterogeneity with language diversity. Second, huge volume of noise is added into the MLOD due to an automatic extraction process or from the very noisy nature of the data source as compared to the monolingual environment [8, 9, 10]. If we consider the case of DBpedia (i.e. the main multilingual source in LOD), the automatic extraction method employed to extract the different language DBpedia edition from Wikipedia by itself contributes to noise. Besides, it is not uncommon to observe quality variations across Wikipedia language (i.e. the source of DBpedia chapters) editions. For instance, there are properties that do not have meaning across languages or there exist missing info box property values. The third challenge is the incompleteness of the data resulting from the open nature of the LOD. In this regard, publishers using two different languages may have different knowledge about an entity, i.e., they may want to see the entity from possibly orthogonal point of views. The fourth challenge comes from the presence of billions of entities; computing the similarity of entities by directly comparing their literal values is computationally expensive and inaccurate. The last challenge comes from the fact that textual descriptions of entities generated from the web of documents are language dependent.

It is assumed that utilizing the structural and textual evidence of entities can handle all the challenges listed in the preceding paragraph. We made RDF graph analysis to discover structural similarities among entities while textual similarity of entities is discovered by projecting into the sub space of the multilingual latent topic space built from a comparable corpus through SVD. Finally, the two independent similarity evidences are combined so as to reach into single global evidence.

The main contributions of this approach are as follows:

MEDDES'15, 25-29 October 2015, Caraguatatuba/Sao Paulo, Brazil

http://dx.doi.org/10.1145/2857218.2857221

- By extracting entities' textual information from the web of documents, we have devised a mechanism of supplementing linked open data entities' descriptions to the problem of entity co-reference resolution. Through this, we have provided a use case where the web of documents can be used together with the linked open data for entity co-reference resolution.
- We have come up with an efficient sparsely promoting tensor factorization algorithm that can scale to DBpedia (i.e. to the number of DBpedia entities and predicates).
- We have adopted a linear opinion pool, an approach used to aggregate probabilistic expert opinion, to combine the structural and textual similarity of entities.
- We have proposed a language independent approach with an initial experiment that demonstrates a comparable result with language dependent approaches to compute textual similarity of entities across languages.

The rest of this article is organized as follows: in section two, we briefly review related works. Section three introduces our proposed method in detail. The description of the data set and experimental set up is presented in section four. The experimental results on French and English DBpedia are reported in section five followed by conclusion and future work in section six.

2. RELATED WORKS

Although the importance of cross-lingual entity linking was stated in several works [32, 33], as far as our knowledge is concerned, an automatic mechanism to link co-referent entity URIs across languages is only proposed in [18,31]. These papers were based on translating language specific descriptions of entities by using machine translation systems. This kind of approach is highly dependent on the quality of the translation system which is barely available in many languages [19].

Entity co-reference resolution is known in database research community with various names such as record linking, duplicate detection or instance identification. However, the RDF data model semantics and the huge number of heterogeneous sources of LOD need further research [3, 6]. Moreover, researches in database have not addressed the cross lingual aspect [18]. In the semantic web, works only focus to link co-referent entity URIs referring to the same real-world object in monolingual data sets [2, 3, 4, 5, 6] or small scale and domain specific RDF datasets [20,21,24, 23, 22].

It is also well-researched in Natural Language Processing with a goal of finding entity mentions within a document or across multiple documents [35,36]. Since, entity co-reference resolution is related to many research areas; it causes confusion about the techniques and solutions relevant to each research area. However, the main dimension one should take into account is the kind of data that are taken into account for the comparison [36]. The techniques and solutions to be applied on unstructured data is not the same to ontology instances [36].

A survey of techniques for entity co-reference resolution is stated in [35,36]. Some techniques apply various attribute-based similarity functions to aggregate the results of value matching methods over the values of properties belonging to two individuals being compared. Others uses transitivity of pre-existing identity links such as owl:sameAs to infer additional equivalence relationship between individuals. Ontological restrictions defined by domain ontologies like owl:FunctionalProperty or owl:InverseFunctionalProperty have been also used to infer new equivalent relationship. In this case, the assumption is equivalent values of inverse functional properties imply the equivalence of the subjects of properties whereas equivalent values of functional properties imply the equivalence of the objects of properties. The other category of techniques is based on identity information available in external knowledge sources. This is a scenario where instances of two data sources are not linked to each other directly but there exist connections to the same external data source. In general, the first technique can be classified as statistical while the remaining belongs to semantic based approach. The proposed technique in this paper combines the semantic and statistical approach, but in a different and novel way. First, entity information contained in the datasets (i.e. structural or relational) being matched and information contained in external sources (i.e. textual descriptions from the web of documents) is utilized. Second, as we are dealing with data sources represented in different natural languages, we used pre-existing identity links and ontological restrictions in the preprocessing stage so as to reduce heterogeneity among data sources. Thirdly, relational learning through RESCAL tensor factorization is used to find the structural similarity of entities. Nickel (2012) introduced RESCAL tensor factorization to the problem of entity co-reference resolution. RESCAL tensor factorization is also applied to capture the joint evidence of entities relationship and literal value descriptions in [7]. The aforementioned studies have portrayed the potential of tensor factorization to find co-referent entities in the domain of LOD. However, they say only little about the possibility of tensor factorization to find co-referent entity identifiers represented in different natural languages. Besides, this research devised a novel method of combining analysis results from tensor models to two way analysis results.

3. CROSS LINGUAL ENTITY CO-REFERENCE RESOLUTION (CLECR)

Before moving directly to the proposed approach, let us first introduce notations used throughout the report. Given the set of URI references U, the set of blank nodes B, and the set of literals L, a triple t := (s, p, o)εG := (UUB)×U×(UUBUL) is called an RDF triple, where s is called subject, p is predicate, and o is object. Given RDF triples, our aim is to identify co-referent entity identifiers located in several LOD sources provided that the LOD sources are described in different natural languages. The proposed approach is shown in figure-1 and detail description is given in the subsequent sections.

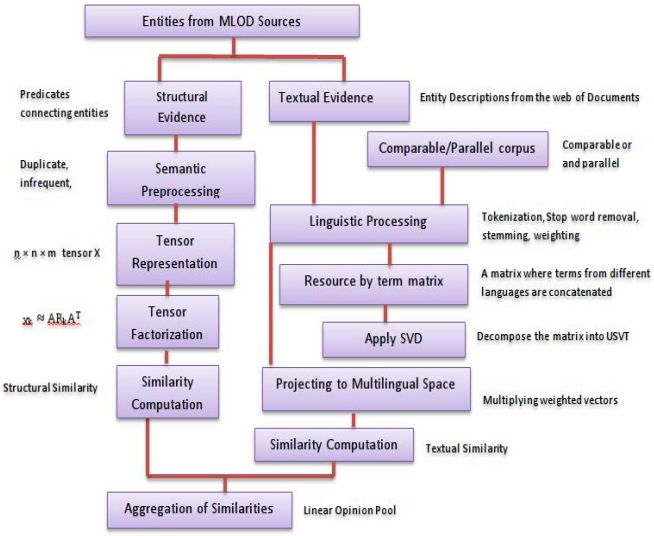

Figure 1. Proposed Approach.

3.1 Structural Evidence

The first evidence to identify co-referent entities comes from the relationship of entities (i.e. entities' structural similarity) with the assumption that entities sharing the same relation to the same entity are co-referent [12]. Third order tensor is used to model the structural evidence of entities. The components of the model are entities and predicates. Since we have not made any distinction between ontological and instance knowledge, entities in this context include all resources, classes and blank nodes while predicates consist of all entity to entity relations.

Nevertheless, there are duplicate predicates and entities, less frequently used predicates, as well as misused predicates. Hence, we used existing explicit equivalent relationship indicating links like owl:equivalentclass, owl:equivalentproperty, owl:SameAs, functional and inverse functional properties to merge duplicate entities or predicates. On the other hand, predicates occurring below a certain threshold have been discarded. Assuming n entities and m predicates, such data is represented as a three-way tensor X of size i × i × k, where the entries on two modes of the tensor correspond to the combined entities appearing both in a subject or object position and the third mode holds the different types of predicates. A tensor entry $X_{ijk} = 1$ implies the k^{th} predicate exists between the i^{th} and j^{th} entity and otherwise the entry becomes $X_{ijk} = 0$.

RESCAL factorization is applied to decompose the three way tensor into a core tensor and a factor matrix. The main reasons to apply tensor factorization are stated as follows. First, the MLOD is obtained through an automatic extraction process so that noise could be introduced either from the process or from the nature of the data source itself. In such settings, tensor factorization has brought an effective result by removing the noisy correlation among features [25,26]. Second, the domain of MLOD is characterized by high dimensionality and data sparcity which adds noise to extract valuable information, efficiently correlate and magnify the latent relationships among various dimensions. Besides, in high dimensional and sparse data, distance measure such as Euclidean distance or cosine similarity has shown odd properties to compute the similarity of entities. However, tensor factorization has brought significant performance improvement in such cases [28,29]. Third, there are lots of empirical evidences in which tensor factorization brought good results. For instance, PARAFAC2 decomposition is used to cluster documents across multiple languages in [16]. Tensor factorization based on Tucker method is also applied to find multiple types relations among words in [29]. It is also applied to identify user behavior for web personalization and recommendation of items in [28]. Tensor factorization is also used to find the relationship of Web pages with words, documents and links in [30]. Lastly, tensors fit to the triple structure of the RDF data model.

RESCAL also scales very well to large number of data sets and has produced high result in various link prediction and entity resolution tasks [7, 11]. Its mathematical properties can better fit to the nature of our dataset. However, as already stated earlier, the three ways tensor constructed from the RDF model has equal number of instances in its first two modes. Therefore, we used RESCAL a special case of DEDICOM (decomposition into directional components) that expect the two modes to be symmetric as a constraint to decompose a three way tensor [34].

Given a tensor X, we compute RESCAL factorization of the tensor into a product of a single matrix A and core tensor R_k as shown in figure-2.

$$X_k \approx AR_kA^T, \text{ for } k = 1,\ldots, m \qquad (1)$$

Where A is an n×r matrix, R_k is a full asymmetric r × r matrix and r is a user-given parameter that specifies the number of latent components or factors. The factorization of the tensor X given in equation (1) is computed by minimizing the following objective function.

$$\min A,R_k \|X_k - AR_kA^T\|_F^2 \qquad (2)$$

An efficient alternative least square algorithm is also proposed to compute equation (2) based on ideas from [34].

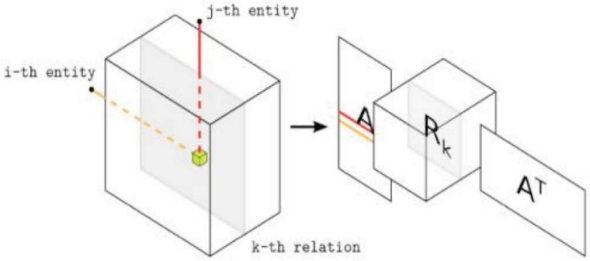

Figure 2. RESCAL Factorization.

The rows of the factor matrix A represent entities in the domain whereas the columns can be considered the discovered latent variables. The entries specify how much an entity participates in a latent space.

Formally, entities participating in the factor matrix A can be represented in r-dimensional vector $e_1,\ldots,e_n$. The vector $e_j = (e_{j,1}, \ldots, e_{j,r})$ represents entity j's participation to the r discovered latent variables, where $e_{j,k}$ is entity j's participation to the k^{th} latent variable.

With the assumption that co-referent entities are more structurally similar than non-co-referent ones, we proposed to compute the similarity of entities from the factor matrix. In order to retrieve entities that are similar to a particular entity e_i with respect to all relations in the data, it is sufficient to compute ranking of entities by their similarity to e_i in A. For instance, given two entities e_i and e_j, their similarity can be computed by equation (3).

$$\text{COS}(e_i,e_j) = \frac{e_i . e_j}{\|e_i\| \|e_j\|} \qquad (3)$$

The cosine similarity result tells us to what extent an entity is similar to others. The higher the cosine similarity value among any given set of entities, the more likely co-referent they are. One of the parameters tried in this experiment is the effect of the number of predicates and the frequency of

predicates used to find co-referent entity representations. Figure-3 is a snap shoot of structural similarity of seven entities to each other computed from the top 20 most frequent predicates. As shown in the figure, the similarity result is inflated. It gives higher similarity value even for entities that are not co-referent. This is a pattern of result when we use few most frequently used predicates only. A reversed pattern is also observed to the similarity result by using all predicates in the data set. An optimal result from the structural evidence is reached by using entities occurring more than 100.

	0	1	2	3	4	5	6
0	1.000000	0.999580	0.967531	0.999810	0.998925	0.999569	0.942464
1	0.999580	1.000000	0.972705	0.998947	0.999254	0.999045	0.946857
2	0.967531	0.972705	1.000000	0.963276	0.977182	0.971056	0.991111
3	0.999810	0.998947	0.963276	1.000000	0.998255	0.999230	0.937804
4	0.998925	0.999254	0.977182	0.998255	1.000000	0.999362	0.955927
5	0.999569	0.999045	0.971056	0.999230	0.999362	1.000000	0.949739
6	0.942464	0.946857	0.991111	0.937804	0.955927	0.949739	1.000000

Figure 3. Structural Similarity Ranking.

However, this similarity result is computed from the structural evidence and cannot be used alone to reach on the co-referent decision. Therefore, we proposed to introduce further textual evidence from the webpages of the web of documents. A discussion of the textual evidence is made in the next section.

3.2 Web of Documents Evidence

Much of the information in the LOD cloud is given as literal values. The number of predicates that connect an entity to a literal – also known as data type properties – are greater than the number of predicates that connect entity to entity (i.e. object type property). This shows entity co-reference resolution systems that do not consider literal information of entities loss significant knowledge about them.

However, there are challenges to use literal information. The first challenge is since there are billions of entities; computing similarity by directly comparing literal values is computationally expensive and inaccurate. Therefore, as many LOD entities are obtained from existing web pages of the web of documents (e.g., DBpedia is converted from Wikipedia), we proposed to extract textual evidence from the web of documents that can supplement entities' descriptions. This could be easily achieved as software tools that identify mentions of LOD entity from the web of documents are matured (e.g., DBpedia Spotlight [13]). Besides, from our own observation, most of the LOD entities have links to descriptions in the form of web pages.

The second challenge comes from the fact that textual descriptions of entities generated from the web of documents are language dependent which requires a similarity computation mechanism across languages. In this study, we used latent semantic analysis (LSA), the standard approach for cross lingual document analysis, to find the similarity of entity descriptions across languages with the use of a parallel or/and comparable corpus [14, 15].

From a parallel or/and comparable corpus, the approach build an entity-by-term matrix X. From this matrix, the row represents entities whereas the columns represent the concatenation of terms from different languages [16]. We then applied SVD to reduce the entity-by-term matrix X into lower dimensional representation as shown in equation (4).

$$X=USV^T \qquad (4)$$

Where U is an entity-by-concept matrix, S is a set of singular values and V is a term-by-concept matrix

Once the LSA model is built, for every entity in A from section 3.1, their textual descriptions from the web pages of the web of documents are extracted. For instance, if an entity comes from the French DBpedia chapter, its textual description could be extracted from French web pages while textual description from English web pages is extracted for entities of English DBpedia. After linguistic preprocessing (viz. removing empty words, stop word removal, stemming and applying a log-entropy weighting scheme), each entity is represented in a vector e_j.

e_j is then projected into a multilingual topic space by multiplying e_j to US^{-1}. The multiplication result of e_j and US^{-1} is e_j's multilingual latent-topic space participation. Formally, the textual evidence of entities can be represented in s-dimensional vector $e_1,...,e_n$. The vector $e_j = (e_{j,1}, \ldots, e_{j,s})$ represents entity j's participation to the s multilingual latent-topic space, where $e_{j,k}$ is entity j's participation to the k^{th} multilingual latent-topic space. In order to retrieve entities that are similar to a particular entity e_i with respect to the textual evidence, it is again sufficient to compute a ranking of entities by their similarity to e_i in A using equation (3) as shown in figure-4. A discussion of the combining strategies are discussed in the following section.

	0	1	2	3	4
0	1.000000e+00	9.679155e-18	-3.445180e-16	-2.566580e-15	-1.540316e-14
1	9.679155e-18	1.000000e+00	-3.544641e-14	-2.633713e-13	-1.612300e-12
2	-3.445180e-16	-3.544641e-14	1.000000e+00	6.136825e-16	8.919308e-15
3	-2.566580e-15	-2.633713e-13	6.136825e-16	1.000000e+00	-1.899449e-15
4	-1.540316e-14	-1.612300e-12	8.919308e-15	-1.899449e-15	1.000000e+00
5	2.335229e-14	2.409014e-12	-2.296879e-14	5.895577e-15	2.174527e-15
6	2.361429e-16	-7.369468e-15	-9.450711e-15	3.520625e-15	8.745646e-15

	5	6
0	2.335229e-14	2.361429e-16
1	2.409014e-12	-7.369468e-15
2	-2.296879e-14	-9.450711e-15
3	5.895577e-15	3.520625e-15
4	2.174527e-15	8.745646e-15
5	1.000000e+00	-2.723406e-15
6	-2.723406e-15	1.000000e+00

Figure 4. Textual Similarity Ranking.

3.3 Merging Evidences

In figure 3 and 4, a screen shot only showing the degree of similarity of an entity e_i to the other 7 entities including to itself is presented. However, in its actual sense, each of the two sources provide the degree of similarity of entity e_i to the other n entities, where n is the number of entities in the domain. This means we are dealing with n mutually exclusive probabilistic outcomes.

Thus, we represent the similarity evidence about entity e_i in an n-dimensional random vector X where each component is itself a random variable, i.e., $X = (X_1, X_2, \ldots, X_n)$. Each of the sources describe the common random vector X having possible outcomes $\{x_i\}^n$, where $n < \infty$ and probability distribution $g_j(x_i) = P_j(X = x_i)$, $j = 1, \ldots, n$ where j is the source.

We applied a linear opinion pool that involves taking a weighted linear average of the sources evidence to combine the two probabilistic similarity vectors into a single vector as shown in equation 5.

$$p(x) = \sum_{j=1}^{n} w_j p_j(x) \qquad (5)$$

Where n is the number of sources (i.e. in this case is 2), $p_j(x)$ represents source j's probability distribution for the random variable x, w_j denotes the weight associated to source j's evidence.

The determination of the weights is a problem which arises when using linear opinion pool. In the literature, a frequently

used strategy is assigning equal weights to each source when there is nothing which suggests that a source's evidence is better than any other one. In our case, the two sources (i.e. the structural and textual) contain complementary information about any given entity. Hence, we give an equal weight to each of the evidences. So the method of combining the two probability vectors into a single vector is reduced to computing their arithmetic average. Lastly, p(x), the aggregated probability vector is used as a global similarity to decide co-referent entities. Hence, for an entity e_i, any entity e_i having the highest probability of similarity from the aggregated vector is considered to be co-referent.

4. EXPERIMENTAL SETUP

The goal of this section is to show to what extent the methods proposed in this research and the various parameters works. All algorithms have been implemented in Python programing language. Besides, the python libraries used includes; numpy, scipy, nltk, pandas, rdflib, SPARQLWrapper, scikit-learn, and sktensor.

4.1 Data Set

The experiment is made with RDF data sets taken from the 2014 French and English localized DBpedia editions. From this data set, entities and predicates of the person class as well as entities belonging to the depth first transitive closure of predicates starting with subjects and objects from the person class are used to conduct the structural analysis. The descriptions of person class data set according to our SPARQL query result is shown in table 4. Please note that the number of triples include triples generated from other classes or data sources generated depth first transitive closure search.

DBpedia edition	No.of entities	Entity-entity relation	Total triples	entity to entity triples
English	2,135,040	3812	1,08550914	71905610
French	208,797	980	27440312	19878087

Table 1. statistics of Data set.

We have then applied many data preprocessing tasks with the extracted data set. In the first preprocessing task, we removed less frequently used predicates with the assumption that exerting much effort to process these predicates is not important as they only bring little performance variation. For instance, from 3812 predicates, only 654 of them occur more than 100 times in the whole data set. Surprisingly, 1376 predicates occur only once and 3107 predicates have below 100 occurrence. We have also observed a similar pattern to predicates of the French DBpedia in which only 385 and 595 predicates occur more than and less than or equal to 100 times respectively. 171 of them also occur only once. In this experiment, predicates occurring less than 100 times are omitted.

In the second data processing stage, we have selected common predicates occurring both in the French and English DBpedia. Identifying predicates which are common to the two sub graphs is straight forward as most predicates are mapped to the DBpedia permanent ontology. For instance, the French predicate http://fr.dbpedia.org/property/nationalité is already mapped to the http://dbpedia.org/property/nationality ontology. Therefore, if http://fr.dbpedia.org/property/nationalité exist in the French sub graph and http://dbpedia.org/property/nationality exist in the English graph, and if they are linked or there exist an equivalent relationships indicating link between them, they are considered to be common predicates. Then, we replace the French predicate by the English predicate in both graphs. In this process, out of 385 predicates from the French DBpedia, 372 are the same to the top 654 English predicates.

Lastly, even within the same sub graph, there exist equivalent predicates. Therefore, using one of them can be sufficient. For instance, http://dbpedia.org/ontology/birthPlace can be used in place of http://dbpedia.org/property/birthPlace as they are connected by the owl:equivalentproperty property. We have conducted reasoning over identity indicating predicates to detect equivalent predicates. By considering predicates connected by owl:equivalentproperty or owl:SameAs, and reasoning through functional and inverse functional properties and their transitive, we have discovered many duplicate predicates. Through this process, the number of predicates in the person class is reduced to 343. From the processed data set, we constructed a tensor X of size 2,750,853x2,750,853x343. RESCAL factorization of the tensor is made. Lastly, for each row in matrix A, we compute its cosine distance to other rows. This similarity ranking is then converted into probability vector and considered as the result of the structural analysis.

To conduct the textual analysis, we extracted textual information from the web of documents in this case Wikipedia. Hence, for entities taken from the English DBpedia, we extract text from the webpages of English Wikipedia, for French DBpedia entities, we extract text from French Wikipedia and for entities generated from other linked open data sources, we have extracted textual descriptions from other sources. Then, natural language processing tasks such as stop word removal and stemming are done. After applying if.idf weighting scheme, each of the entities are represented as a weighted vectors.

We train our cross lingual latent topic model by using 40,290 highly comparable Wikipedia articles extracted and aligned at article level from the French and English Wikipedia. Further description of this dataset is available in [17]. From this dataset, a 40,290x640519 resource by term matrix is built. This matrix is subjected to SVD so as to reduce it to lower dimensional space. Once the model is constructed, the weighted entity vectors derived from Wikipedia are multiplied by the US^{-1}.

4.2 Test data set

Since the French DBpedia sub graph has already been linked to the English version, we used the existing owl:sameAs links as a ground truth to evaluate the performance of the proposed approach. Hence, we purposely selected 200 entities from the English DBpedia edition that is connected by owl:sameAs ontology to the French DBpedia so as to evaluate the approach. Two things are considered while selecting the test data set. The first is entities that participate in building the cross lingual latent topic model are not included in the test data set. The second is the test samples are purposely distributed to subclass of the person class. In other words, we selected 40 entities from each sub class in the person class (athlete, actor, artist, politician and scientist).

5. RESULTS AND DISCUSSION

The main challenge we face in this research is lack of bench mark data set. This problem is also reported in [18]. The OAEI2010 and OAEI2011 data set, a commonly used bench marking data set to test the performance of entity co-reference resolution systems in semantic web, has no cross lingual coverage. Therefore, we evaluated our proposed approach by using the test data set explained in section 4.2. This evaluation strategy is also used in [31]. By using the test data set, our approach (i.e.CLECR) has an average precision of 95% as

shown in figure-5. Both qualitative and quantitative comparison of the proposed system is also made with two other systems.

In [31], a system which we named as Tatiana Lesnikova evaluate the suitability of a Machine Translation to interlink RDF resources described in English and Chinese languages and found a precision of 98%. Another system named Relational in [7] reported an average precision of 97% on benchmark datasets of OAEI 2010 and 2011. The first system uses machine translation which is not available in many languages or costly to develop while the latter is only tested in a monolingual setting. The result is promising as we are trying to achieve a challenging goal using language independent approach with cheaply available comparable corpora. Besides, the performance difference between CLECR and the others is statistically insignificant.

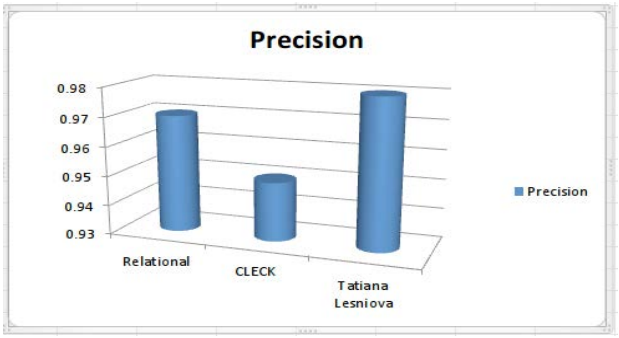

Figure 5. Performance Comparison with the State of Art Systems.

Furthermore, another experiment is conducted to show the effect of combining the textual and structural evidences as compared to using these sources independently. The result in figure-6 is an average precision of the various parameters tested in this experiment. As shown in the figure, combining the textual and structural evidence brings high performance. The average performance of the structural evidence is better as compared to the textual evidence. However, the structural evidence revealed performance irregularities with different parameters such as on the number and type of predicates used. For instance, we observed a performance degradation tendency when we only use most commonly used predicates, and exhaustively use all predicates including very few infrequent ones. To the contrary, the performance of the textual evidence remains consistent across parameters.

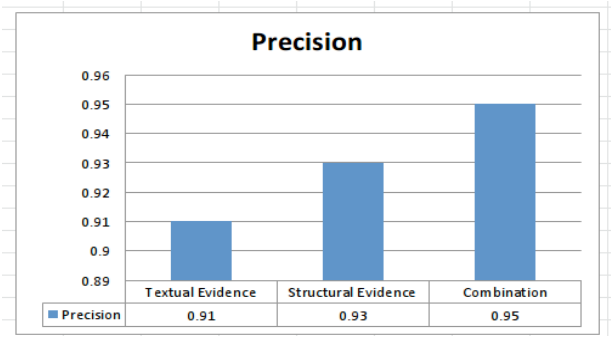

Figure 6. Performance across Evidences.

6. CONCLUSION AND FUTURE WORK

This paper proposed a mechanism to identify co-referent entities across languages located in several LOD sources. We showed the significance of tensor factorization in detecting co-referent entities across languages. We have come up with an efficient RESCAL factorization algorithm that can scale to large number of entities and predicates. By combining the structural and textual information of entities, we provided a use case where the web of document can support and/or can be used together with the web of linked open data. Besides, we have also adopted a linear opinion pool to the task of aggregating probabilistic similarity evidences for a given entity. Experiments on selected data sets from the English and French DBpedia evaluate the contribution of the proposed approach to find co-referent entity references across languages.

To aggregate the structural and textual evidence, we assume equal weight to both sources that reduce the computation to linear averaging. Nevertheless, in all cases taking the linear average of the two evidences may negatively affect the global similarity result. Hence, we need further research to investigate ways of weight optimization while combining similarity evidences from the two sources. Moreover, we also plan to increase the vocabulary coverage of the parallel/comparable corpus which has significant effect to the overall performance. Finally, further inquiry will also be made so as to let the tensor decomposition algorithm scalable to the LOD scale.

7. REFERENCE

[1] Böhm, C., de Melo, G., Naumann, F., & Weikum, G. (2012, October). LINDA: distributed web-of-data-scale entity matching. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2104-2108). ACM.

[2] Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., & Decker, S. (2012). Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Web Semantics: Science, Services and Agents on the World Wide Web, 10, 76-110.

[3] Hu, W., Chen, J., & Qu, Y. (2011, March). A self-training approach for resolving object coreference on the semantic web. In Proceedings of the 20th international conference on World wide web (pp. 87-96). ACM.

[4] Böhm, C., de Melo, G., Naumann, F., & Weikum, G. (2012, October). LINDA: distributed web-of-data-scale entity matching. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2104-2108). ACM.

[5] Suchanek, F. M., Abiteboul, S., & Senellart, P. (2011). Paris: Probabilistic alignment of relations, instances, and schema. Proceedings of the VLDB Endowment, 5(3), 157-168.

[6] Song, D., & Heflin, J. (2011). Automatically generating data linkages using a domain-independent candidate selection approach. In The Semantic Web–ISWC 2011 (pp. 649-664). Springer Berlin Heidelberg.

[7] de Assis Costa, G., & de Oliveira, J. M. P. A Relational Learning Approach for Collective Entity Resolution in the Web of Data.

[8] Kim, E. K., Weidl, M., Choi, K. S., & Auer, S. (2010). Towards a Korean DBpedia and an Approach for Complementing the Korean Wikipedia based on DBpedia. In OKCon (pp. 12-21).

[9] Al-Feel, H. (2013). A Step towards the Arabic DBpedia. International Journal of Computer Applications, 80(3), 27-33.

[10] Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., & Metakides, G. (2012). Internationalization of linked data: The case of the greek

dbpedia edition. Web Semantics: Science, Services and Agents on the World Wide Web, 15, 51-61.

[11] Nickel, M., Tresp, V., & Kriegel, H. P. (2012, April). Factorizing YAGO: scalable machine learning for linked data. In Proceedings of the 21st international conference on World Wide Web (pp. 271-280). ACM.

[12] Singla, P., & Domingos, P. (2006, December). Entity resolution with markov logic. In Data Mining, 2006. ICDM'06. Sixth International Conference on (pp. 572-582). IEEE.

[13] Men Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (pp. 1-8). ACM.

[14] Kim, W., & Khudanpur, S. (2004, May). Cross-lingual latent semantic analysis for language modeling. In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on (Vol. 1, pp. I-257). IEEE. [15] Zhang, D., Mei, Q., & Zhai, C. (2010, July). Cross-lingual latent topic extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1128-1137). Association for Computational Linguistics.

[16] Chew, P. A., Bader, B. W., Kolda, T. G., & Abdelali, A. (2007, August). Cross-language information retrieval using PARAFAC2. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 143-152). ACM.

[17] Smith, J. R., Quirk, C., & Toutanova, K. (2010, June). Extracting parallel sentences from comparable corpora using document level alignment. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 403-411). Association for Computational Linguistics.

[18] Lesnikova, T. (2013). NLP for interlinking multilingual LOD. In Proc. ISWC Doctoral consortium (pp. 32-39). No commercial editor..

[19] Fu, B., Brennan, R., & O'Sullivan, D. (2009). Cross-lingual ontology mapping–an investigation of the impact of machine translation. In The Semantic Web (pp. 1-15). Springer Berlin Heidelberg.

[20] Nikolov, A., Uren, V., Motta, E., & De Roeck, A. (2008). Integration of semantically annotated data by the KnoFuss architecture. In Knowledge Engineering: Practice and Patterns (pp. 265-274). Springer Berlin Heidelberg.

[21] Noessner, J., Niepert, M., Meilicke, C., & Stuckenschmidt, H. (2010). Leveraging terminological structure for object reconciliation. In The Semantic Web: Research and Applications (pp. 334-348). Springer Berlin Heidelberg.

[22] Jentzsch, A., Zhao, J., Hassanzadeh, O., Cheung, K. H., Samwald, M., & Andersson, B. (2009, September). Linking Open Drug Data. In I-SEMANTICS.

[23] Shi, L., Berrueta, D., Fernández, S., Polo, L., Fernández, S., & Asturias, A. (2008, October). Smushing RDF instances: are Alice and Bob the same open source developer. In PICKME Workshop.

[24] Sleeman, J., & Finin, T. (2010, November). Learning co-reference relations for FOAF instances. In 9th International Semantic Web Conference (ISWC2010).

[25] Skillicorn, D. (2007). Understanding complex datasets: data mining with matrix decompositions. CRC press.

[26] Nickel, M. (2013). Tensor factorization for relational learning (Doctoral dissertation, lmu).

[27] Carvalho, A., & Larson, K. (2013, August). A consensual linear opinion pool. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (pp. 2518-2524). AAAI Press.

[28] Rawat, R. (2010). User behaviour modelling in a multi-dimensional environment for personalization and recommendation.

[29] Chang, K. W., Yih, W. T., & Meek, C. (2013). Multi-Relational Latent Semantic Analysis. In EMNLP (pp. 1602-1612).

[30] Kolda, Tamara G., Brett W. Bader, and Joseph P. Kenny. "Higher-order web link analysis using multilinear algebra." Data Mining, Fifth IEEE International Conference on. IEEE,2005.

[31] Lesnikova, T., David, J., & Euzenat, J. (2014). Interlinking English and Chinese RDF Data Sets Using Machine Translation. In Proc. 3rd ESWC workshop on Knowledge discovery and data mining meets linked open data (Know@ LOD), Hersounisos (GR) (Vol. 2013).

[32] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web, 11, 63-71.

[33] Buitelaar, P., Choi, K. S., Cimiano, P., & Hovy, E. H. (2012). The Multilingual Semantic Web (Dagstuhl Seminar 12362). Dagstuhl Reports, 2(9), 15-94.

[34] PHAN, A. H., & CICHOCKI, A. Tensor Decompositions for Very Large Scale Problems.

[35] Beheshti, S. M. R., Venugopal, S., Ryu, S. H., Benatallah, B., & Wang, W. (2013). Big data and cross-document coreference resolution: Current state and future opportunities. arXiv preprint arXiv:1311.3987.

[36] Scharffe, F., Fan, Z., Ferrara, A., Khrouf, H., & Nikolov, A. (2011). Methods for automated dataset interlinking.