Cross-lingual link discovery with TR-ESA

Fedelucio Narducci[a], Matteo Palmonari[b], Giovanni Semeraro[a]

[a] *Department of Computer Science, University of Bari Aldo Moro*
Via E. Orabona 4, I-70125 Bari, Italy

[b] *Department of Informatics, Systems and Communication, University of Milano Bicocca*
Viale Sarca, 336, I-20126 Milano, Italy

Abstract

Cross-lingual data linking is the problem of establishing links between *resources*, such as places, services, or movies, which are described in different languages. In cross-lingual data linking it is often the case that very short descriptions have to be matched, which makes the problem even more challenging. This work presents a method named TRanslation-based Explicit Semantic Analysis (TR-ESA) to represent and match short textual descriptions available in different languages. TR-ESA translates short descriptions in any given language into a pivot language by exploiting a machine translation tool. Then, it generates a Wikipedia-based representation of the translated text by using the Explicit Semantic Analysis technique. The resulting representations are used to match short descriptions in different languages. The method is incorporated in CroSeR (Cross-lingual Service Retrieval), an interactive data linking tool that recommends potential matches to users. We compared results coming from an in-vitro evaluation on a gold standard consisting of five datasets in different languages, with an in-vivo experiment that involved human experts supported by CroSeR. The in-vivo evaluation confirmed the results of the in-vitro evaluation and the overall effectiveness of the proposed method.

Keywords: Cross-lingual Matching, Cross-lingual Data Linking, Wikipedia

Email addresses: fedelucio.narducci@uniba.it (Fedelucio Narducci), palmonari@disco.unimib.it (Matteo Palmonari), giovanni.semeraro@uniba.it (Giovanni Semeraro)

1. Introduction and Motivations

The Linked Data paradigm has been proposed to publish structured data on the web in a way that data can be easily consumed by third-party applications [18]. Several tools can be used to transform data into Resource Description Framework (RDF)[1], a format compliant to the Linked Data principles. However, publishing an RDF dataset on the web is not sufficient to realize the vision of linked data. To interconnect two datasets, a *data linking* task has to be performed. The task consists in discovering and representing links between resources described in two datasets. These resources are usually instances in a knowledge base such as persons, places, movies, services, and so on. Links specify relations between resources using an unambiguous semantics. For example, the *owl*:*sameAs*[2] property represents a link between two resources that denote a same real-world entity; the *skos*:*narrowMatch* and *skos*:*broadMatch* properties, defined in the SKOS[3] vocabulary, represent the relations between two resources such that the first one is more general than the second one and viceversa, respectively. Cross-lingual data linking is defined as the problem of establishing links between resources described in different languages. It is emerging as a new research topic because of the rapid growth of the multilingual web of data [16, 36]. As of November 2015, more than 1,000,000 Open Government Datasets have been published online by national and local governments from more than 40 countries in 24 different languages[4]. Cross-lingual data linking tasks, which are complicated because of language and socio-cultural barriers (e.g., two resources may describe services that can be considered equivalent in their respective countries, but they are not exactly the same service), are even more difficult when resources in the two datasets are associated with limited descriptions, e.g., consisting of few words only.

Linking resources described with a small number of words is a hard task because only few words can be used to discover potential matches, introducing a *coverage problem*: only a small number of links can be found with the help of automatic matching methods. Machine translation tools can be used to bridge the language gap so as to deal with many different languages in a scalable way [26, 27]. However, plain translation of descriptions does not mitigate the coverage problem, making word-based similarity measures

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/owl-features/
[3] http://www.w3.org/TR/skos-reference/
[4] http://logd.tw.rpi.edu/iogds_data_analytics

(proposed where richer descriptions are available) ineffective in this context.

The main contribution of this paper is the proposal of a matching function for the cross-lingual data linking between resources with poor textual descriptions. The effectiveness of the proposed matching function and of the interactive data linking approach is evaluated using in-vitro and in-vivo experiments which aim to answer to the following research questions:

1. **R1**: Is it possible to combine in a unique approach the capability to deal with several languages in a scalable way with the capability to enrich input descriptions so as to solve the coverage problem?
2. **R2**: Is it possible to improve the coverage of a cross-lingual matching function by preserving performance in terms of ranking quality?
3. **R3**: Is a coverage-oriented matching function able to solve a real-world cross-lingual data linking task?

To answer to question R1, we propose an unsupervised cross-lingual matching function that combines machine translation and semantic enrichment of textual descriptions. Short textual descriptions in languages other than English are automatically translated into English and enriched using Wikipedia. These enriched descriptions are used to compute the similarity between resources. In particular, we describe in detail an enrichment method that exploits the Explicit Semantic Analysis (ESA) technique [14] to build representations consisting of vectors of Wikipedia concepts. The application of ESA to translated descriptions, referred as TRanslation-based Explicit Semantic Analysis (TR-ESA) in the rest of the paper, introduces a novel cross-lingual matching function explicitly targeted to maximise coverage. An in-vitro experiment conducted on six different languages evaluate different matching functions based on machine translation by considering the lists of ranked results that they return for a set of input resources. Experiments show that TR-ESA achieves much better coverage, measured by Hit Rate [8], than matching functions that either do not enrich the descriptions or perform the enrichment steps by means of different state-of-the-art methods.

To answer to question R2, we also evaluate (in the in-vitro experiment) accuracy and ranking quality of the matching functions based on machine translation using two well-known measures: accuracy@n and Mean Reciprocal Rank [52]. Experiments show that TR-ESA achieves higher accuracy (for every language and for every $n > 1$) than any other matching function because it can retrieve correct links for a much larger number of resources. While increased accuracy and coverage come at the price of lower ranking

quality, correct links retrieved by TR-ESA fall on average between the fourth and fifth position. In other words, experiments suggest that, in difficult cross-lingual matching tasks, a matching function that significantly increases coverage may achieve better accuracy with limited loss in terms of ranking quality.

To answer to question R3, we integrated TR-ESA in an interactive linking tool. The tool is a web application that recommends a ranked list of potential matches, given an input resource. Users can browse the list of recommended matches and establish a link by choosing among three different relations, i.e., *owl:sameAs*, *skos:broadMatch*, *skos:narrowMatch*. In an in-vivo experiment, we asked to 15 domain experts to use our tool to perform a new cross-data linking task between two datasets consisting respectively of 750 services described in Italian and 1435 services described in English. With the help of the tool, users have been able to discover and establish links for 452 Italian services.

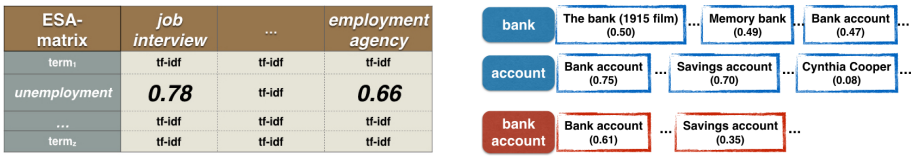
The rest of this paper is organized as follows. Section 2 describes TR-ESA and the methodology used to obtain Wikipedia-based representations of short textual descriptions. Section 3 describes our link discovery approach. Section 4 presents CroSeR, the system that implements our approach. In Section 5, experimental results are presented and, in Section 6, we compare our work with previous work on cross-lingual matching carried out in different research fields. Finally, in Section 7, we discuss conclusions and future work.

2. A Semantic Matching Function for Short Textual Descriptions

Matching two or more texts is essential for several artificial-intelligence tasks, such as classification, clustering, filtering, and retrieval. Text matching can be implemented as simple string matching, which analyzes the lexical overlap between two texts, or can take into account also their semantics.

In this section, we present a semantic-based matching function able to deal with short textual content in different languages. The data linking strategy adopted in the paper is based on this matching function, whose goal is to enrich short textual content. This function is at the same time, language agnostic, and thus potentially exploitable for any language. Shortness of the text and multilinguality are two characteristics that make the matching task more challenging.

In order to cope with shortness of the text, we exploited Explicit Semantic Analysis (ESA) [14], which allows to represents terms and documents using Wikipedia concepts. In order to cope with multilinguality, ESA has



(a) *(b)*

Figure 1: (a) The ESA*-matrix; (b) an example of semantic interpretation vectors*

been properly modified to handle short texts written in any language. For this purpose, we defined a translation-based version of ESA (TR-ESA) that extends the application of the technique to any language for which a machine translation tool is available.

2.1. Explicit Semantic Analysis

ESA uses Wikipedia as a space of semantic concepts explicitly defined and described by humans. Formally, given the space of Wikipedia concepts $C = \{c_1, c_2, ..., c_k\}$, a term t_i can be represented by its *semantic interpretation vector* $v_i =< w_{i1}, w_{i2}, ..., w_{ik} >$, where w_{ij} denotes the strength of the association between t_i and c_j, with $1 \leq j \leq k$. Weights are computed and stored into a matrix T, called ESA-*matrix*, in which each of the k columns corresponds to a concept, and each row corresponds to a term of the Wikipedia vocabulary (i.e., the set of distinct terms in the corpus of all Wikipedia articles). The item $T[i, j]$ contains w_{ij}, the TF-IDF value of term t_i in the article (concept) c_j. Therefore, the semantic interpretation vector for a given term is the corresponding row vector in the ESA-*matrix* (Figure 1a). As an example, the meaning of the term *unemployment* can be described by a list of concepts (the semantic interpretation vector) it refers to, the Wikipedia articles for: *job interview, employment agency, ...* The semantic interpretation vector for a text fragment f (i.e. a sentence, a document, a resource description) is obtained by computing the centroid (average vector) of the semantic interpretation vectors associated with terms occurring in f.

The motivation behind the decision of using ESA in our matching function is twofold:

- ESA is able to perform a sort of word sense disambiguation (WSD) based on the semantics *explicitly* associated to the target term by humans [14];
- ESA is able to generate *new* knowledge in terms of the Wikipedia concepts most related to the input text fragment; this process is also referred to as *feature generation*.

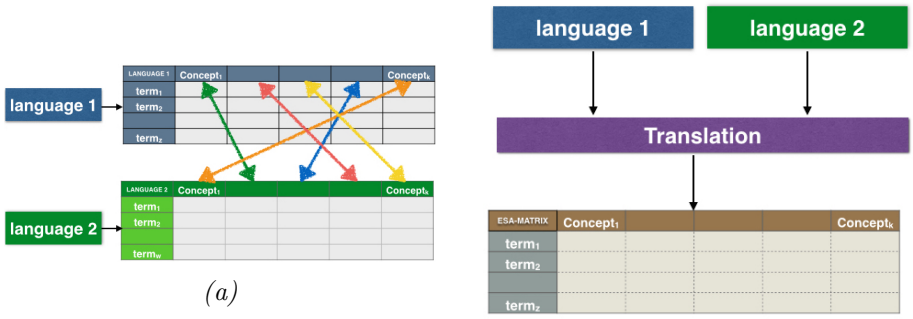


Figure 2: Example of CL-ESA *(a) and* TR-ESA *(b) with two languages*

We expect that these two characteristics can be particularly useful in a cross-lingual matching scenario, when the textual descriptions to match can be the output of - possibly imprecise - machine translation services.

Let us consider an example with the short resource description *bank account*. The term *bank* is a polysemous word, with meanings related to finance, geography, computing, etc. If we extract the semantic interpretation vector (Figure 1b) for *bank*=⟨The bank (1915 film) (0.50), Memory bank (0.49), ..., Bank account (0.47), ...⟩ and *account*=⟨Bank account (0.75), Savings account (0.70), ..., Cynthia Cooper (accountant)(0.08), ...⟩, then the computation of their centroid vector results in the semantic interpretation vector for *bank account*: ⟨Bank Account (0.61), Savings Account (0.35), ...⟩. It is worth noting that the most related concept in that specific context occurs in the first position, thus realizing a form of WSD. Concerning the second motivation behind the adoption of ESA, we can consider the resource description *Home Schooling*. ESA generates the following centroid vector: ⟨Home (0.67), School (0.55), Education (0.48), Family (0.35), ...⟩. The vector adds new knowledge that does not explicitly appear in the input text, thus realizing a form of semantic enrichment of the short textual description.

2.2. TR-ESA: a new cross-lingual version of Explicit Semantic Analysis

The first adaptation of ESA to cross-lingual scenarios has been proposed by Sorg and Cimiano [50], who named their approach Cross-Lingual Explicit Semantic Analysis (CL-ESA). The main idea behind CL-ESA is the same one behind the ESA model. However, this model is able to generate a uniform Wikipedia-based representation for texts in different languages. Given an input text, the semantic interpretation vector is built, similarly to ESA. The main difference is that the space C is defined as the intersection between the

sets of concepts of the different Wikipedia versions the system deals with. More formally, $C_{cl-esa} = C_{l_1} \cap C_{l_2} \cap ... \cap C_{l_n}$, where C_{l_i} is the set of Wikipedia concepts for the language l_i. Therefore, for each language l_i the ESA-matrix has to be built (Figure 2a). Since the set C_{cl-esa} is the intersection between the concepts in different Wikipedia versions, the cardinality of the smallest C_{l_i} is the upper bound of the cardinality of C_{cl-esa}. Accordingly, in the case the system handles languages with a poor Wikipedia version, the small cardinality of C_{cl-esa} may represent a weakness for the feature generation process. For this reason, when at first we decided to exploit CL-ESA in our system for a language for which Wikipedia has limited coverage, we clashed with very disappointing results.

In order to overcome these limitations we based our matching function on a new original variant of ESA that we named TRanslation-based Explicit Semantic Analysis (TR-ESA). TR-ESA implements the ESA model with a translation step performed prior to the feature generation process (Figure 2b). TR-ESA has the advantage of building only one ESA-matrix (from the most accurate language - i.e. English), since the input text is preliminarily translated in the same language of the matrix. The idea behind TR-ESA is quite simple, nonetheless we show that is effective for cross-lingual tasks. First, an ESA-matrix is built for English, the language for which the richest Wikipedia version is available. A text in any language for which a machine translation tool is available is translated into English. Then, the translated text is enriched by using the English ESA-matrix. In this way, it is possible to perform the feature generation process for any language for which a machine translation tool is available and/or an accurate Wikipedia version is not available.

TR-ESA is not affected by the completeness of the Wikipedia versions of the languages the system deals with. Indeed, TR-ESA generally uses the richest Wikipedia version, that is to say, the English one.

After the feature generation step, in order to compute the similarity between two texts we represent them by vectors in a multidimensional space in which each dimension is a Wikipedia concept. Formally, each resource is represented as a vector $\vec{r} =< w_1, \ldots, w_n >$, where w_k is the TF-IDF value of the Wikipedia concept k. It is also possible to exploit other weighting schemas, such as the simpler boolean vector. Finally, the similarity between two resources (vectors) s and s' is computed in terms of cosine similarity, which measures the cosine of the angle between the two vectors as follows:

$$cosSim(s, s') = \frac{\sum_{i=1}^{n} s_i * s'_i}{\sqrt{\sum_{i=1}^{n} (s_i)^2} * \sqrt{\sum_{i=1}^{n} (s'_i)^2}} \quad (1)$$

3. Interactive Cross-Lingual Data Linking

TR-ESA is used as a feature generation and matching method in the interactive approach to cross-lingual data linking [16, 26, 27] proposed in the paper.

Definition 1. *(Cross-lingual data linking) Let S and T be two sets of resources, called source (S) and target (T) dataset, described in two different languages L_1 and L_2 respectively. Let R be set of relations between resources in S and T. A cross-lingual data linking task can be defined as a partial function $l : S \times T \rightarrow R$, defined as follows:*

$$l(s_i, t_j) = r_w, \tag{2}$$

where:

- $s_i \in S$ *is a resource described in* S*;*
- $t_j \in T$ *is a resource described in* T*;*
- $r_w \in R$ *is a relation occurring between* s_i *and* t_j*.*

Given two datasets describing respectively n and m resources, a manual approach to link the two datasets would require $n * m$ comparisons, which makes manual linking of data often unfeasible even for datasets of medium size (e.g., with hundreds of resources). Conversely, an interactive data linking approach would help users by reducing the number of comparisons they have to perform in order to establish links.

More precisely, given a resource s_i, an interactive data linking system helps users to reduce the target set T to a subset of resources T_{s_i} which are estimated as candidate matches with respect to one or more relations in R. Therefore, the user can avoid the burden to scan the whole set T for finding the right resource to be connected to s_i, by limiting the analysis to the subset T_{s_i} (much smaller than T) of the most promising resources. This reduction is performed by computing a semantic similarity score between s_i and all the resources in T so as to present a ranked list of candidate matches to the user. Users can exploit the returned list in order to validate the links between the source s_i and the candidate target resources in T_{s_i}.

Concerning the set of relations R, the relation most frequently used in data linking is *owl:sameAs*, which establishes that two resources denote the same real-world object. However, it has been observed that *owl:sameAs* has often been used with different meanings, or even misused, in linked

data [17]. One problem to consider when linking instances is that they may represent strongly related objects, described at different levels of granularity [24], which can be considered the same under certain perspectives. For example, one resource may denote a localized version of a movie, e.g., the distribution of a movie for the Italian market, which is dubbed in Italian, while another resource denotes the master representation of the movie.

These observations lead us to consider in particular two further relations taken from the SKOS vocabulary [31]: *skos:broadMatch* and *skos:narrowMatch*. These relations establish links between two resource descriptions (SKOS concepts are reified as ontology instances) such that the first one is more specific or more general, respectively, than the second one. The three relations - *owl:sameAs*, *skos:broadMatch*, *skos:narrowMatch* - cover the spectrum of possible relations (equivalent, more specific, more general) between resources possibly described at different levels of abstractions. The relations *skos:broadMatch* and *skos:narrowMatch* are specifically introduced in SKOS to represent matches between elements of different knowledge systems and are domain-independent. Of course, these relations can be specialized in a given domain. However, at present, we preferred properties of widely adopted vocabularies, such as OWL and SKOS, to domain specific or linguistic-based properties, with the goal of maximizing interoperability even at the price of using a generic terminology.

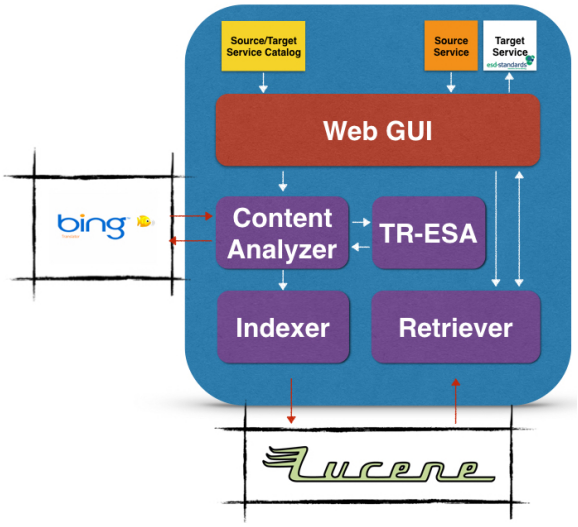


Figure 3: General architecture of the cross-lingual link discovery system

The general architecture of the system supporting the interactive cross-lingual data linking approach based on TR-ESA is depicted in Figure 3. Our approach is based on the hypothesis that semantic annotations of resource descriptions can support the linking task in an effective way, even when the resources are poorly described in a not uniform way. Experimental results discussed in Section 5.1 will confirm this hypothesis. The system consists of two main components: the *Content Analyzer* and the *Retriever*. The *Content Analyzer* processes the resource descriptions and builds a semantic annotation for each resource by exploiting *TR-ESA*; the *Retriever* generates a set of resources that could be linked to an input resource by a relation $r \in R$ by exploiting the Lucene[5] index managed by the *Indexer*. The user interacts with the system through the *Web GUI*.

Content Analyzer. Resources described in different languages are the input to the *Content Analyzer*. Before generating a Wikipedia-based representation, a machine translation process, powered by Bing[6], is performed. More specifically, each *resource description* is translated into English. Subsequently, translated descriptions are used by another component called TR-ESA that is able to generate an ESA-based representation of the resources. Therefore, for each resource s, a set of Wikipedia concepts W_s semantically related to the resource description is generated; we call this set W_s *Wikipedia-based annotation* of s. Wikipedia-based representations are then indexed by exploiting the *Indexer* component.

Wikipedia-based annotations aim to capture the main topics related to a resource. Furthermore, the annotation of a resource with a set of Wikipedia concepts could represent an additional link between the resource and the Linked Open Data cloud by using a connection hub like DBpedia.

Retriever. Indexed resources are represented by using the Vector Space Model (VSM). A multidimensional space in which each dimension is a Wikipedia concept is thus built. Accordingly, a resource is a point in that space. The *Retriever* computes the cosine similarity between a vector representing the source item and a vector representing the target item, and generates a ranked list of resources related to the source item.

4. CroSeR for Cross-lingual Linking of E-gov Services

The cross-lingual link discovery approach described in Section 3 was implemented in a system named CroSeR (Cross-lingual Service Retrieval).

[5]https://lucene.apache.org/

[6]http://www.microsoft.com/en-us/translator/

CroSeR supports users in the specific task of linking e-gov services described in different languages. In this domain, S represents the source service catalog, T is the target service catalog, and R is the set of relations defined as $R = \{owl{:}sameAs,\ skos{:}narrowMatch,\ skos{:}broadMatch\}$. The target service T is the European Local Government Service List (LGSL).

LGSL, as part of the Electronic Service Delivery (esd)-toolkit website[7], is one of the main interesting results of the SmartCities project[8]. The SmartCities project involves seven countries of the North Sea region: England, Netherlands, Belgium, Germany, Scotland, Sweden, and Norway. Each country is responsible to build and maintain its list of public services delivered to the citizens, and all those services are interlinked to the services delivered by other countries. In addition, LGSL is already linked to the LOD cloud. The goal of LGSL is to build standard lists that define the semantics of public sector services. Services in LGSL describe abstract functionalities of services that are concretely offered by a number of providers at a local level. A LGSL service such as *Homelessness support* represents a category of services, rather than an individual service. However, following an approach also used by other e-gov service representation models, these categories are represented in a knowledge base as instances and can be referred to as *abstract services* [42]. For this reason, two services that are considered equivalent by domain experts and that belong to different catalogs in different languages are linked through a *owl:sameAs* link.

Even if the service catalogs in the ESD-toolkit have been linked using the *owl:sameAs* relation (established by human experts), several linked services may still represent services described at different levels of abstraction. For example, in LGSL the Dutch service *Kwijtschelding belastingen* (*Remission of tax* in English) has been linked to the (certainly more specific) service *Council tax discount*. This kind of mistakes may be motivated by the need to make easy the data linking tasks for the human experts, who were not supported by a tool like CroSeR when they had to manually link their catalogs to the LGSL (which consists of 1,435 distinct services).

The aim of CroSeR is therefore to support the discovery of links between a new e-gov service catalog and LGSL, according to the semantics adopted in the esd-toolkit. Automatic cross-lingual matching methods, which can reduce the effort needed to manually link these catalogs, have to deal with the poor quality of the service descriptions. Services are represented by

[7] http://www.esd.org.uk
[8] http://www.smartcities.info/

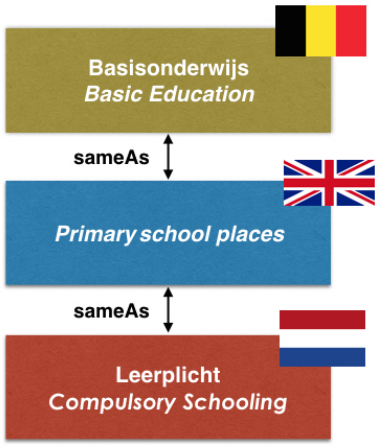


Figure 4: Examples of linked services in the LGSL. *The machine translation into English powered by Bing is reported in italic.*

minimal descriptions that often consist of the name of the service and very few data.

4.1. Local Government Service List and Linked Open Data

By looking at the examples of links established by domain experts to the LGSL catalog shown in Figure 4, we can see that the labels associated with linked services are not a mere translation from a language to another. As an example, the Belgian service *Basisonderwijs* (literally translated as *Basic Education*) and the Dutch service *Leerplicht* (literally translated as *Compulsory Schooling*) have been manually linked to the English service *Primary school places* by domain experts. Therefore, the automatic matching of the text labels associated to services is not a trivial task, which shows how a semantic matching approach like TR-ESA can be beneficial to cross-lingual data linking.

4.2. CroSeR@work

In a typical use case, the user first uploads the service catalog - a set of services in a structured form - into the system (this functionality is disabled in the demo version available online[9]). After that, the catalog is semantically

[9]http://dacena.disco.unimib.it:8080/croser/

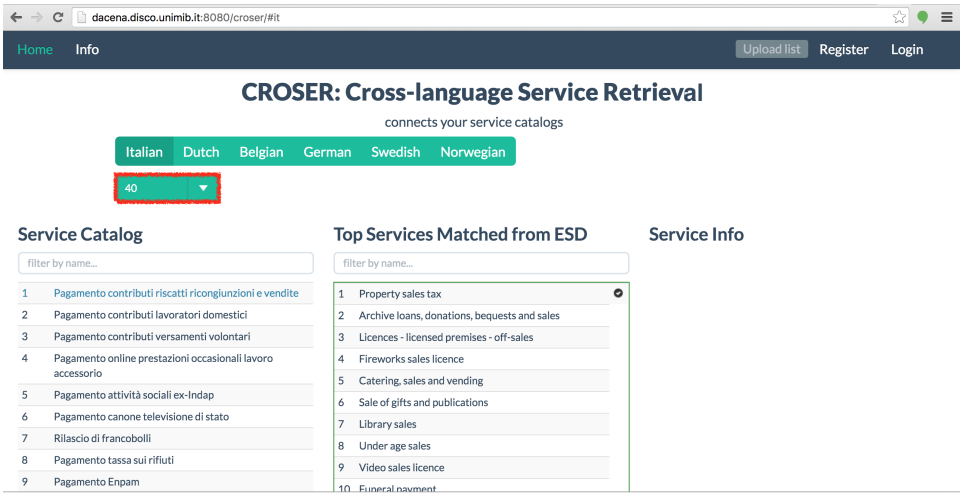


Figure 5: Selection of the number of candidate services

analyzed and indexed. The user is now able to explore the catalog by scrolling the whole list of services or by performing a keyword-based search.

Next, the user selects a *source* service from the catalog and CroSeR retrieves a list of *candidate target services* from LGSL that are potentially linkable by a *skos:narrowMatch*, *skos:broadMatch*, *owl:sameas* predicate. The number of candidate services to retrieve can be configured by the user (Figure 5).

Finally, users can switch on the feedback mode of CroSeR (registration and login are required), thus the system stores the relation between the source service and the LGSL service after selecting a candidate service from the retrieved list.

The matching can be also performed by a simple keyword-based search. Keyword-based matching turns out to be particularly useful when the user encountered a service that is eligible for a relation with the current source service, but that service is not suggested in the first positions by CroSeR.

CroSeR currently supports seven languages: English, Italian, Dutch, Belgian, German, Swedish, and Norwegian. However, it can be easily extended to other languages by configuring the Bing translation services.

Finally, some services in the list of candidate services are marked by a check. These services have been manually linked by human experts and represent the gold standard that we used in our in-vitro experiments (see Section 5.1). The user can compare the suggestions generated by the system with the human annotations. CroSeR can also support users in revising

already existing links[10].

5. Experimental Evaluation

We carried out two experimental sessions: an *in-vitro* experiment useful to detect the best system configuration, and an *in-vivo* experiment in which CroSeR was exploited for helping human experts to link an Italian catalog of e-gov services to the LGSL.

We tested our approach in the e-gov domain for different reasons:

- First, linking public services descriptions is a real-world problem of interest for many governments involved in Open Data initiatives. Linking public services is an objective of the European Community to support the integration of services across countries[11] and an increasing number of initiatives for publishing and interlinking data about public services have been undertaken. In this context, LGSL has been created and published as Linked Open Data. In another Italian project, named SMART, a public service list has been created and published as Linked Open Data so as to support a variety of applications, from personalization of service compositions [1] to strategic planning [42].

- Second, in the SmartCities project, domain experts manually linked services described in different languages to services described in English (i.e., LGSL). These links have been used as a gold standard to evaluate the performance of our matching methods on a variety of languages (Dutch, Belgian, German, Swedish, and Norwegian) and datasets[12]. In addition, in the context of the SMART project, we conducted an in-vivo experiments by having domain experts using the CroSeR application to establish new links from a large Italian service catalog to the LGSL catalog. Indeed, our methodology was exploited to support human experts to link the Italian Public Administration Services (IPAS) catalog of e-gov services to LGSL.

- Third, e-gov service descriptions encompasses several domain specific vocabularies (e.g., education, culture, transport, sport, etc.). This aspect gives generality to the experimental evaluation.

[10] A demo video of CroSeR is available at
https://www.dropbox.com/s/c4uxhy7p23cahus/Croser demo.mp4?dl=0

[11] http://ec.europa.eu/digital-agenda/en/towards-cloud-public-services

[12] We observe that the authors of this paper did not take any part in this project or in the linking tasks performed therein. The gold standards used in the experiments discussed in this paper have been independently created by third parties prior to this work.

From a preliminary analysis, we have noticed that the IPAS and the LGSL catalogs contain services described at different levels of abstraction, as other catalogs already linked in the ESD-toolkit. This confirmed the intuition that, during the user study, the user should have been allowed to define also relations different from the *owl:sameAs* statement as defined in ESD, namely *skos:narrowMatch*, and *skos:broadMatch*. In this way we could evaluate the capability of the system also in suggesting candidate services for different kinds of relations.

5.1. Session 1: in-vitro experiment

The *in-vitro* evaluation is carried out on five catalogs already linked to the LGSL, used as gold standards. Links between services belonging to different catalogs are in terms of *owl:sameAs* statements established by human experts. The goals of this experiment are: (1) to compare the effectiveness of different service representations, (2) to evaluate the capability of the system in boosting the correct service in the first positions of the ranked list, (3) to evaluate the capability of the system to face the coverage problem (as defined in Section 1). The first goal is evaluated by means of the accuracy@n metric, the second goal by means of the Mean Reciprocal Rank metric, and the third goal by means of the Hit Rate. More details are reported in the next section.

5.1.1. Design and datasets

The five gold standards used in this session consist of *owl:sameAs* links between the Dutch, German, Belgian, Swedish, and Norwegian service catalogs and the English catalog, i.e., LGSL. These six datasets and the links are extracted from the ESD-toolkit catalogues freely available online[13]. The catalogues have been created independently by public bodies in the respective countries and then manually linked. The authors of this paper have not taken any part neither in the creation of the catalogs nor in the linking process. Each catalog contains a different number of services, each of which associated with a label of variable length. Some statistics are shown in Table 1. It is worth noting that, even if Dutch and Belgian services are represented in the same language (i.e, Dutch), services often have different labels. For example, the English service *Primary school places* has the label *Leerplicht* in the Dutch catalog, whereas it has the label *Basisonderwijs* in the Belgian one.

[13]http://standards.esd-toolkit.eu/EuOverview.aspx

Table 1: Dataset distribution

Language	# **services**	**average length**
English (EN)	1435	3.24
Belgian (BE)	341	1.94
German (DE)	190	2.87
Dutch (NL)	225	2.63
Norwegian (NO)	165	2.98
Swedish (SV)	66	2.59
		2.71

We indexed every catalogue used in the experimental evaluation (English, Dutch, German, Belgian, Swedish, and Norwegian). For each service, we extracted, translated and represented its textual label in terms of Wikipedia concepts using TR-ESA.

The labels have an average length of about three words.

To evaluate our approach we compare performance against several alternative lexical matching methods that can easily scale to handle a large number of languages because they are based on machine translation. In every method, original descriptions are first translated using the Bing APIs into English. It is worth to note that we exploited Bing since its translation service is available for free. In our experiment we compare the performance of TR-ESA (see Section 2), with the performance of

1. a baseline method that does not provide any semantic analysis step;
2. several alternative methods that, by leveraging state-of-the-art *entity linking* tools, produce Wikipedia-based representations, thus performing a sort of semantic analysis.

The baseline method considered in our experiment is keyword-based matching. For a keyword-based representation, only stemming and stopword elimination are performed on the text.

Alternative methods used in our experiments exploit Wikipedia-based representations extracted using Wikipedia Miner, Tagme, DBpedia Spotlight, Babelfy. All these methods perform a disambiguation process in order to assign the correct Wikipedia concepts to the input text. We report a brief description of each system.

Wikipedia Miner. Wikipedia Miner is a tool for automatically cross-referencing documents with Wikipedia [32]. The software is trained on

Wikipedia articles, and thus learns to disambiguate and detect links in the same way as Wikipedia editors [12].

Tagme. Tagme is a system that performs accurate and on-the-fly semantic annotation of short texts via Wikipedia as knowledge base [13]. The annotation process is composed of two main phases: the *anchor disambiguation* and the *anchor pruning*. This process takes into account the probability of the anchor text to be used as link in Wikipedia and the coherence between the candidate page and the candidate pages of other anchors in the text.

DBpedia Spotlight. DBpedia Spotlight [30] was designed with the explicit goal of connecting unstructured text to the LOD cloud by using DBpedia as hub. Also in this case, the output is a set of Wikipedia articles related to a text retrieved by following the URI of the DBpedia instances.

Babelfy. Babelfy is a novel graph-based approach to link text fragments to BabelNet synsets [33]. BabelNet synsets identifies word senses and/or named entities described in multiple languages and linked to Wikipedia articles. Babelfy uses BabelNet 1.1.1 [37]. The main advantage of Babelfy is a unified approach to solve the two tasks of Entity Linking and Word Sense Disambiguation in any of the languages covered by the *native* multilingual semantic network.

All the previously mentioned entity linking systems are on-line services. They take a text description (the service label) as input, and return a set of *Wikipedia concepts* that emerge from the input text. All those services allow to configure some parameters in order to favor recall or precision. Given the conciseness of the input text in our domain, we set those parameters for improving the recall instead of precision. English-translated descriptions are processed by using those tools in order to find Wikipedia entries that are relevant to the input text. Wikipedia entries are used to generate two types of representations: purely *concept-based representations*, such that every document is represented by a vector of Wikipedia entries, and *hybrid representations* obtained by merging the keywords extracted from the service label with the concept-based representations. Figure 2 shows the different annotations generated by the systems for the service label *Home schooling*. The *hybrid representation* adds to each annotation the keywords the label is composed of. For example, in the case of *tagme*, the *tagme+keyword* representation will be *{Home Schooling, Homeschooling}*.

We use three different metrics in our experiments: *accuracy@n* (*a@n*), *Mean Reciprocal Rank* (MRR) [52], and *Hit Rate*. The *a@n* is calculated considering only the first n retrieved services. If the correct service occurs in the *top-n* resources, the service is marked as correctly retrieved. We considered different values for n, with $n = 1, 3, 5, 10, 20, 30$. The second

Table 2: Example of Wikipedia annotations generated by different systems for the service label Home schooling

Approach	Annotations
Wikipedia Miner	[[Homeschooling|Home schooling]]
Tagme	en.wikipedia.org/wiki/Homeschooling
DBpedia Spotlight	dbpedia.org/resource/Home_Schooling
Babelfy	bn:00000356n, bn:00562314n, bn:00093341v
TR-ESA	Home, School, Education, Family, [...]

metric (MRR) considers the rank of the correctly retrieved service and is defined as follows:

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}, \tag{3}$$

where $rank_i$ is the rank of the correctly retrieved service ($service_i$) in the ranked list, and N is total number of services in the catalog. The higher the position of the services correctly retrieved in the list is, the higher the MRR value for a given representation is. The *Hit Rate* indicates the percentage of services in the catalog for which the system is able to suggest the correct link.

5.1.2. Results

Table 3 and Table 4 report results of *accuracy@n*, *MRR* and *Hit Rate* for each language. As regards accuracy@n the best configurations are the TR-ESA-based ones for all $n > 1$. The more the number of retrieved services is (i.e., n), the higher the gap with respect to the baseline is. When TR-ESA is compared to keyword-based configuration we observe an average improvement of +21% for $n = 10$, +33% for $n = 20$, and +41% for $n = 30$.

TR-ESA is the only representation that shows no improvements when Wikipedia concepts are combined with keywords (i.e., *tr-esa+keyword*). This is due to the fact that TR-ESA generally outperforms the keyword-based representation, thus adding keywords produces no benefits, as expected.

The worst performance is shown by Wikipedia Miner, followed by Tagme and Babelfy. This is likely due to difficulties in detecting concepts in really short text fragments. Indeed, those representations are really accurate when they are able to identify Wikipedia concepts into the processed text. Differently from TR-ESA, those representations improve their accuracy by merging Wikipedia concepts with keywords, but they generally do not outperform

the representation only based on keywords (except for *dbpedia+keyword, and babelfy+keyword*, which show a slight improvement). In terms of MRR, the representation with the best values is Wikipedia Miner followed by Babelfy, and Tagme (which are the representations with the worst performance in terms of $a@n$). However, since CroSeR is a retrieval system, and not, for example, a question-answering engine (for which to have the correct answer in the first position plays a crucial role), we prefer representations that obtain a good accuracy with an acceptable MRR value. In this context, TR-ESA represents the best compromise between accuracy and ranking. Indeed, the average rank of the correct service for the TR-ESA representation is between the $fourth$ and the $fifth$ position of the retrieved list[14]. Therefore, we consider the results satisfying.

The best performance for *Hit Rate* is obtained by TR-ESA-based configurations in every language. That is an interesting outcome since it demonstrates that even though representations obtained with TR-ESA are more noisy than representations enriched with Tagme, Wikiminer, Babelfy, and DBpedia Spotlight (as showed in Subsection 5.3), TR-ESA leads to a vast improvement in terms of Hit Rate (up to +65%, +62% on average) with respect to the baseline, yet maintaining satisfying levels of accuracy and ranking.

There are also differences in terms of minimum and maximum accuracy values (Table 5) among the different catalogs and representations (all minimum values are obtained by a@1, and all maximum values by a@30). The highest MIN value is obtained by TR-ESA followed by the keyword baseline. The highest MAX value is obtained by TR-ESA, as well. However, in this case TR-ESA is followed by *tagme+keyword*, *dbpedia+keyword*, and *wikiminer+keyword*, and *babelfy+keyword* . All those representations show a large gap with respect to the 'pure' representation composed only of Wikipedia concepts. We observe that the best and the worst results are achieved with different languages. For example the highest value is achieved on the Norwegian catalog (NO) by TR-ESA and this confirms the high accuracy of our methodology also with a language for which a poor Wikipedia version is available. The lowest result is on the Swedish (SV) catalog by Tagme. These differences may be explained on the ground of the different performance shown by machine translation tools on different languages.

Results obtained in this in-vitro experiment suggest that TR-ESA is the most effective representation. In order to statistically validate these results,

[14]RR is 1 if the first relevant document is retrieved at rank 1, it is 0.5 if the first relevant document is retrieved at rank 2 and so on.

Table 3: Results for the Dutch, Belgian, and German languages. The highest values are reported in bold.

Dutch								
Configuration	**@1**	**@3**	**@5**	**@10**	**@20**	**@30**	**MRR**	**Hit Rate**
keyword	0.33	0.46	0.50	0.54	0.54	0.55	0.61	48.24%
tagme	0.12	0.17	0.18	0.18	0.19	0.19	0.64	16.47%
tagme+keyword	0.32	0.45	0.48	0.55	0.56	0.57	0.56	50.20%
wikiminer	0.08	0.09	0.11	0.11	0.11	0.11	**0.75**	9.41%
wikiminer+keyword	0.32	0.44	0.48	0.53	0.54	0.55	0.59	48.24%
tr-esa	0.31	**0.48**	**0.54**	**0.62**	**0.69**	**0.72**	0.38	**72.55%**
tr-esa+keyword	0.31	**0.48**	**0.54**	**0.62**	**0.69**	**0.72**	0.38	**72.55%**
dbpedia	0.18	0.24	0.24	0.25	0.26	0.26	0.70	22.75%
dbpedia+keyword	0.33	0.45	0.50	0.56	0.57	0.57	0.57	50.59%
babelfy	0.14	0.16	0.17	0.18	0.19	0.19	0.76	18.67%
babelfy+keyword	**0.34**	0.46	0.49	0.55	0.56	0.56	0.60	56.44%
Belgian								
Configuration	**@1**	**@3**	**@5**	**@10**	**@20**	**@30**	**MRR**	**Hit Rate**
keyword	0.32	0.43	0.48	0.54	0.56	0.56	0.57	56.30%
tagme	0.09	0.13	0.14	0.14	0.14	0.14	0.65	14.37%
tagme+keyword	**0.33**	0.44	0.52	0.57	0.58	0.58	0.57	58.36%
wikiminer	0.11	0.14	0.14	0.15	0.15	0.15	**0.77**	14.96%
wikiminer+keyword	**0.33**	0.46	0.49	0.56	0.58	0.58	0.56	57.77%
tr-esa	**0.33**	**0.48**	0.53	**0.61**	**0.68**	**0.71**	0.39	**84.16%**
tr-esa+keyword	**0.33**	**0.48**	**0.54**	**0.61**	**0.68**	**0.71**	0.39	**84.16%**
dbpedia	0.20	0.25	0.26	0.28	0.29	0.29	0.70	29.03%
dbpedia+keyword	**0.33**	0.45	0.52	0.54	0.58	0.58	0.57	58.36%
babelfy	0.10	0.13	0.14	0.14	0.14	0.14	0.67	14.37%
babelfy+keyword	**0.33**	0.44	0.52	0.57	0.58	0.58	0.57	58.36%
German								
Configuration	**@1**	**@3**	**@5**	**@10**	**@20**	**@30**	**MRR**	**Hit Rate**
keyword	0.24	0.40	0.47	0.49	0.49	0.49	0.49	49.47%
tagme	0.15	0.17	0.18	0.18	0.18	0.18	0.82	17.89%
tagme+keyword	0.26	0.41	0.47	0.51	0.51	0.51	0.51	51.05%
wikiminer	0.26	0.41	0.47	0.50	0.51	0.51	0.51	50.53%
wikiminer+keyword	0.22	0.34	0.40	0.42	0.43	0.43	0.51	50.53%
tr-esa	**0.29**	**0.43**	**0.51**	**0.57**	**0.66**	**0.71**	0.35	**82.63%**
tr-esa+keyword	**0.29**	**0.43**	**0.51**	**0.57**	**0.66**	**0.71**	0.35	**82.63%**
dbpedia	0.16	0.20	0.22	0.22	0.22	0.22	0.76	21.58%
dbpedia+keyword	0.27	**0.43**	0.49	0.52	0.52	0.52	0.53	52.11%
babelfy	0.15	0.17	0.18	0.18	0.18	0.18	**0.82**	21.66%
babelfy+keyword	0.26	0.41	0.47	0.51	0.51	0.51	0.51	51.05%

Table 4: Results for the Norwegian, and Swedish languages and Average values of accuracy@n, MRR, and Hit Rate. The highest values are reported in bold.

Norwegian								
Configuration	**@1**	**@3**	**@5**	**@10**	**@20**	**@30**	**MRR**	**Hit Rate**
keyword	0.27	0.37	0.44	0.49	0.53	0.53	0.51	53.33%
tagme	0.12	0.16	0.17	0.18	0.18	0.18	**0.67**	18.18%
tagme+keyword	0.27	0.41	0.48	0.52	0.56	0.56	0.48	56.36%
wikiminer	0.11	0.12	0.14	0.14	0.14	0.14	0.78	13.94%
wikiminer+keyword	**0.29**	0.39	0.46	0.50	0.55	0.55	0.53	55.15%
tr-esa	0.26	**0.41**	**0.49**	**0.59**	**0.72**	**0.78**	0.30	**88.48**%
tr-esa+keyword	0.26	**0.41**	**0.49**	**0.59**	**0.72**	**0.78**	0.30	**88.48**%
dbpedia	0.16	0.22	0.23	0.26	0.26	0.26	0.63	26.06%
dbpedia+keyword	**0.29**	0.41	0.46	0.50	0.55	0.55	0.52	55.15%
babelfy	0.12	0.16	0.17	0.18	0.18	0.18	0.67	18.18%
babelfy+keyword	0.27	**0.41**	0.48	0.51	0.55	0.55	0.50	55.15%
Swedish								
Configuration	**@1**	**@3**	**@5**	**@10**	**@20**	**@30**	**MRR**	**Hit Rate**
keyword	0.18	0.27	0.32	0.33	0.33	0.33	0.55	33.33%
tagme	0.09	0.15	0.15	0.15	0.15	0.15	0.60	15.15%
tagme+keyword	0.20	0.30	0.33	0.35	0.35	0.35	0.57	34.85%
wikiminer	0.11	0.11	0.11	0.11	0.11	0.11	**1,00**	10.61%
wikiminer+keyword	0.17	0.27	0.32	0.33	0.33	0.33	0.50	33.33%
tr-esa	**0.24**	**0.32**	**0.36**	**0.47**	**0.56**	**0.59**	0.39	**62.12**%
tr-esa+keyword	**0.24**	**0.32**	**0.36**	**0.47**	**0.56**	**0.59**	0.39	**62.12**%
dbpedia	0.09	0.18	0.20	0.20	0.20	0.20	0.46	19.70%
dbpedia+keyword	0.21	0.30	0.33	0.33	0.33	0.33	0.64	33.33%
babelfy	0.09	0.15	0.15	0.15	0.15	0.15	0.60	15.15%
babelfy+keyword	0.20	**0.32**	0.35	0.35	0.35	0.35	0.57	34.85%
Average values								
Configuration	**@1**	**@3**	**@5**	**@10**	**@20**	**@30**	**MRR**	**Hit Rate**
keyword (baseline)	0.27	0.39	0.44	0.48	0.49	0.49	0.54	49.42%
tagme	0.11	0.16	0.16	0.17	0.17	0.17	0.68	16.85%
tagme+keyword	0.28	0.40	0.46	0.50	0.51	0.52	0.54	51.50%
wikiminer	0.11	0.12	0.13	0.13	0.13	0.13	**0.81**	13.09%
wikiminer+keyword	0.27	0.39	0.44	0.48	0.50	0.50	0.54	50.29%
tr-esa	**0.29**	**0.42**	**0.49**	**0.57**	**0.66**	**0.70**	0.36	**79.92**%
tr-esa+keyword	**0.29**	**0.42**	**0.49**	**0.57**	**0.66**	**0.70**	0.36	**79.92**%
dbpedia	0.16	0.22	0.23	0.24	0.24	0.24	0.65	24.43%
dbpedia+keyword	**0.29**	0.41	0.46	0.49	0.51	0.51	0.56	51.26%
babelfy	0.12	0.15	0.16	0.17	0.17	0.17	0.71	17.61%
babelfy+keyword	0.28	0.41	0.46	0.50	0.51	0.51	0.55	51.17%

Table 5: Minimum and Maximum accuracy@n values for each representation

Representation	MIN	Lang	MAX	Lang
keyword	0.18	SV	0.56	BE
tagme	0.09	SV	0.19	NL
tagme+keyword	0.20	SV	0.58	BE
wikiminer	0.08	NL	0.15	BE
wikiminer+keyword	0.17	SV	0.58	BE
tr-esa	**0.29**	DE	**0.78**	NO
tr-esa+keyword	**0.29**	DE	**0.78**	NO
dbpedia	0.09	SV	0.29	BE
dbpedia+keyword	0.21	SV	0.58	BE
babelfy	0.09	SV	0.19	NL
babelfy+keyword	0.20	SV	0.58	BE

we compared the keyword-based representation with TR-ESA for each catalog. Table 6 reports levels of significance obtained by performing the Wilcoxon Matched Pairs Test. More specifically, the number reported in each cell shows the statistic significance (*p-value*) of the differences in a@n (for each n value) between keywords and TR-ESA. Empty cells show no statistically significant differences. We can observe that, for the Belgian catalog (which is also the richest one), the improvement of TR-ESA is statistically significant for each value of n. Conversely, other catalogs show statistically significant differences from $n = 10$ onwards. These results can be considered satisfying, since a set of 10 resources is quite small with respect to the whole LGSL catalog composed of more than $1,400$ services (Google, for example, shows 10 resources in its first page of search results).

Table 6: Results of the Wilcoxon test used to compare Keyword-based representation with TR-ESA*-based one*

Catalog	**a@1**	**a@3**	**a@5**	**a@10**	**a@20**	**a@30**
Dutch				0.01	0.01	0.01
Belgian	0.05	0.01	0.01	0.01	0.01	0.01
German				0.01	0.01	0.01
Norwegian				0.01	0.01	0.01
Swedish				0.01	0.01	0.01

5.2. Session 2: in-vivo experiment

The goal of the in-vivo experiment was twofold: 1) to assess whether results of the in-vitro experiment are confirmed; 2) to evaluate the effectiveness of CroSeR in suggesting candidate services also for relations different from *owl:sameAs*.

5.2.1. Design and dataset

We asked 15 domain experts, who were previously involved in the SMART project, to link IPAS to LGSL by using CroSeR. These users formerly contributed to the definition of the Italian service catalog, thus they were familiar with the services to link. Their background is specifically focused on e-Government and PA service delivery and they have mid-level skill in computer science (final users of several software applications). For each IPAS service we extracted, translated and represented its textual label in terms of Wikipedia concepts by using TR-ESA (the representation that performed best in the in-vitro evaluation).

A set of 50 services from IPAS was randomly assigned to each registered user. In addition, each user was provided with the following instructions/guidelines:

- register to the system and log in;

- select the Italian service catalog;

- select the full list of candidate services;

*- select one of the 50 services from the source service catalog (*IPAS*)*;

- evaluate the list of candidate services;

- if the label of a candidate service is not clear, then click on the URI *to get more details from the* ESD *website;*

- after identifying the correct candidate service, choose one of the following relations: narrower (*skos:narrowMatch*) - the candidate service in LGSL is more specific than the Italian (source) service; broader (*skos:broadMatch*) - the candidate service in LGSL is more generic than the Italian (source) service; sameAs (*owl:sameAs*) - the candidate service in LGSL is equivalent to the Italian (source) service;

- select another service from the Italian list until all 50 source services have been evaluated.

For each established link, we track the position of the matched service in the ranked list returned by CroSeR. In the case the correct service is not located in the first 40 positions of the list of candidate services, we suggest to the users that they can use the full-text search implemented by CroSeR to avoid scrolling very large lists. In this case, users can retrieve even candidate services that are hidden in faraway positions. Finally, a source service could

have a relation with one or more candidate services, but we suggested to define the relation with the first correctly found service. We decided to adopt this simplification in order to avoid cognitive overload to the users and reassure them that, establishing at most one (correct) link for each source service, was sufficient in relation to the experiment goals.

To evaluate CroSeR, we report several details about the experiment and we evaluate the accuracy of the link recommendations by computing two metrics used for the in-vitro evaluation, i.e., *accuracy@n* (*a@n*) and *Mean Reciprocal Rank* (MRR), against the links established by human experts.

5.2.2. Results

Some statistics about the user study are reported in Table 7. The total number of services in the Italian catalog IPAS that were evaluated by users is 750. Users succeeded to link 452 of them to LGSL. The distribution of the defined relations is the following: 77 are of type *skos:narrowMatch*, 301 are of type *skos:broadMatch*, and 74 are of type *owl:sameAs*. This partition confirms our hypothesis that resources may identify objects described at different levels of abstraction. The problem of linking resources that describe similar objects at different levels of abstraction may be particularly relevant in a cross-lingual context, where socio-cultural differences are mirrored by linguistic constructs.

Finally, we evaluated the usage of full-text search for defining a relation. Specifically, full-text search was a suggestion given to users during the experiment, namely to use the full-text search if the correct candidate service did not appear in the first 40 positions of the ranked list. For example, a link to a candidate service that was ranked in the 200th position is considered retrieved by the full-text search. Only 25 relations ($\sim 5,5\%$ of the total number) were defined by using the full-text search, showing the effectiveness of the ranked list generated by CroSeR.

Table 7: Statistics about the in-vivo experiment

# **source services evaluated from IPAS**	750
# **total established links**	452
# **narrower links**	77
# **broader links**	301
# **sameAs links**	74
# **links defined by full-text search**	25

We compared the results obtained in the user study to the results ob-

tained in the *in-vitro* evaluation with the LGSL gold standard and using TR-ESA. Links between services in the LGSL gold standard are only represented in terms of *owl:sameAs* relations, as above mentioned, and are established by human experts of corresponding local governments.

Table 8: Comparison of accuracy@n between in-vitro and in-vivo experiments

Catalog	**a@1**	**a@3**	**a@5**	**a@10**	**a@20**	**a@30**
Italian (broader)	0.196	0.419	0.485	0.665	0.794	0.847
Italian (narrower)	**0.351**	**0.520**	**0.610**	**0.779**	**0.844**	0.909
Italian (sameas)	0.230	0.392	0.541	0.689	0.824	**0.919**
Dutch	0.311	0.480	0.538	0.622	0.689	0.716
Belgium	0.326	0.478	0.534	0.613	0.677	0.707
German	0.326	0.478	0.534	0.613	0.677	0.707
Norwegian	0.261	0.406	0.491	0.588	0.715	0.782
Swedish	0.242	0.318	0.364	0.470	0.561	0.591

The values of *accuracy@n* are reported in Table 8. Generally speaking, the best results for all values of n are obtained in the in-vivo experiment on the Italian catalog. More specifically, best accuracy has been obtained for the *narrower* link on the Italian catalog. For the *sameAs* relation on the Italian catalog, the greater the n value is, the greater the gap is with respect to the other catalogs. The *broader* link obtains good results in terms of *a@n*. Also in this case, accuracy is high for $n >= 20$. Better results on matching *narrower* relations, compared to *broader* ones, is likely due to different levels of abstraction of IPAS services with respect to LGSL ones. Indeed, the different level of abstraction between the source service and target service conditions also the type of relation to be established.

Therefore, one of the main outcomes of the *in-vivo* evaluation is that CroSeR guarantees that $\sim$ 84% (a@30=0.8472) of the correct services are in the first 30 positions, even in the worst cases. This is a good result with respect to the size of LGSL, which consists of 1,435 distinct services. As a consequence, user efforts are drastically reduced by using CroSeR.

The second analysis ocused on the capability of CroSeR to boost relevant services in the first positions of the ranked list of candidate services. Results in terms of MRR for each representation are reported in Table 9. The highest MRR scores are obtained in the in-vivo experiment, for every considered link relation, with the best MRR value being equal to 0.483 for *narrower* link.

Table 9: MRR *values for each catalog*

Catalog	MRR
Italian (narrower)	**0.483**
Italian (broader)	0.348
Italian (sameas)	0.370
Dutch	0.311
Belgian	0.326
German	0.289
Norwegian	0.261
Swedish	0.242

5.3. Discussion

We carried out two distinct experimental sessions: an in-vitro evaluation and an in-vivo experiment with real users. The in-vivo experiment confirmed and consolidated the results obtained by the in-vitro evaluation. The in-vivo experiment showed that the TR-ESA-based matching function implemented in CroSeR can effectively reduce the effort required to human experts to perform their cross-lingual data linking task. The application has helped users to discover links between an Italian e-gov service catalog and the English LGSL catalog, in a scenario where the only other alternative would have been to explore the whole LGSL.

As opposed to other Wikipedia-based representations obtained by means of entity linking tools (i.e., Tagme, Wikipedia Miner, and DBpedia Spotlight), TR-ESA is able to generate semantic representations for a very large set of services. Furthermore, CroSeR helped the users discover services that a full-text search could not find, effectively addressing the coverage problem. As an example, the Italian service *Rilascio certificato di usabilità dei sepolcri* (machine translated into English as *Issuing certificate of usability of graves*) shares no keywords with other related services in the LGSL. Nevertheless, CroSeR was able to suggest the service *Cemeteries and crematoria* in the 31st position, which surely has strong semantic correlation with the source service, despite the fact that it shares no keywords with it. Even though the service does not have a high rank, without CroSeR the user would hardly had success in linking the source service unless she had surfed the whole LGSL. Other similar examples of non trivial matches found by CroSeR are: *Postal license* linked (broader) to *Rilascio francobolli* (machine translated as *Stamp release*), which consists in the capability of a postal office (or similar offices/shops) to sell stamps, and ranked in the 2nd position; *Litigation*

support linked (narrower) to *Arbitrati e conciliazioni* (translated as *Arbitrations and conciliations*), which represents means for solving litigations, and ranked in the 1st position; *Economic information and analysis* linked to *Consultazione studi di mercato* (translated as *Consulting market research*), which represents a service that supplies market studies, and ranked in the 19th position, and so on.

However, we also noticed that the feature generation process based on TR-ESA frequently introduces considerable noise in the textual descriptions. Specifically, some Wikipedia articles were introduced as new features, even though they are not properly related to the service label. Nevertheless, when a textual description is very short, and possibly inaccurate, TR-ESA-based feature generation may still find a matching. Indeed, in our model, it is sufficient that two descriptions of semantically related resources share just a Wikipedia concept to obtain a non zero matching score, even if they share no keywords. For this reason we believe that our approach is very valuable for linking poorly described resources. As an example, we consider the service *Disdetta abbonamento RAI* (machine translated as *Subscription cancellation RAI*), where *RAI* is the Italian national public broadcasting company. TR-ESA generates some noisy concepts like *aishwarya_rai, chiang_rai_city, gilles_de_rais, chiang_rai_province*, but generates also some features that are related to the television domain (*rai_1, rai_2, rai_3, silvio_berlusconi, television*) and that are exploited when matching the service with the LGSL service *Domestic TV and radio license*, which does not share any keywords with the translated Italian service label.

Concerning the 298 services for which users could not find any LGSL service to link, we guess that in most cases there were no services eligible for a relation. However, we found that, for few services, matching failed because TR-ESA provides limited support for expanding acronyms. Indeed, acronyms are processed only if they occur as terms in the ESA-matrix, which can happen for some acronym (see, e.g., the previous example with *RAI*), but not for many of them. For example, the service *Pagamento COSAP*, where *COSAP* is an Italian acronym that means *Canone per l'Occupazione di Spazi ed Aree Pubbliche* (translated in English as *Public space usage fee*) should have the service *Commercial use of municipal land* as a candidate for a narrower relation. However, if the acronym is not expanded, TR-ESA cannot return any match.

Now we can answer to the research questions formulated in Section 1:

1. **R1**: Is it possible to combine in a unique approach the capability to deal with several languages in a scalable way with the capability to

enrich input descriptions so as to solve the coverage problem?
Yes. TR-ESA demonstrated to be effective in dealing with different languages even when resource descriptions are short.

2. **R2**: Is it possible to improve the coverage of a cross-lingual matching function without penalizing performance in terms of ranking quality?
Yes. TR-ESA demonstrated to be the approach achieving best coverage. Nonetheless, its accuracy is comparable and often better than other approaches based on entity-linking algorithms that show higher precision.
3. **R3**: Is a coverage-oriented matching function able to solve a real-world cross-lingual data linking task?
Yes. The integration of TR-ESA in CROSER demonstrated to be effective in suggesting not only *owl:sameAs* relations, but also *skos:narrowMatch* and *skos:broadMatch* relations.

6. Related Work

To better scope the problem addressed in this paper, we report the distinction between multi-language information access (MLIA) and cross-lingual information access (CLIA) proposed in the literature. MLIA is the problem of accessing, querying and retrieving information from collections in any language and at any level of specificity [44]. In this sense, MLIA subsumes CLIA, which is the problem of accessing a data collection in a target language L' by using a source language L, where $L \neq L'$. These definitions can be specialized for multi-language and cross-lingual retrieval, data linking, or matching.

Cross-lingual information retrieval is the problem of finding a set of documents lexicalized in one language that are most relevant to a query lexicalized in a different language [50]. Data linking is the problem of establishing semantic links, i.e., relations associated with a shared semantics, between resources, (usually) ontology instances, described in different data sources (e.g., two places, two movies, two job titles, and so on) [39].

We proposed a cross-lingual matching function for short textual descriptions and an interactive data linking method that uses this function to support users in establishing links between resources. The matching function considers only textual descriptions, which can be assimilated to documents. Thus, our approach is largely inspired by cross-lingual information retrieval, which is the first research field we will compare our work with in subsection 6.1. Afterwards, we compare our work to related work in the emerging field

of cross-lingual data linking in subsection 6.2 [36, 26, 27]. Ontology matching [48] is another problem that can be also related to data linking, and thus can turn out to be relevant for comparison (subsection 6.3). However, ontologies have key features that play an important role in matching such as concepts related in a subconcept graph, which are not usually considered in data linking and in our work. Entity linking is known as the problem of linking mentions of named entities occurring in text to a resource described in a knowledge base [33]. We will compare our work to work done in the field of cross-lingual entity linking in subsection 6.4, despite significant differences exist between this problem and the one addressed in our paper. Finally, the cross-lingual matching problem has been recently investigated also for addressing other tasks such as content-based recommendation [34].

This paper extends the work presented in [35, 36**?**] along several dimensions. First, we formalize and define the TR-ESA approach as a general matching method that can be exploited in different domains and scenarios. Second, we present an implemented system that allows users to perform their data linking task using two new relations in addition to *sameAs*, which was the only relation considered in the previous paper. The adoption of three relations to establish links between real-world objects described at different levels of abstraction can be useful also in other scenarios and may limit the abuse of the *sameAs* relation, as discussed in [17]. Third, in this paper we significantly extend the evaluation by quantifying coverage, reporting detailed results for every language, adding a better qualitative analysis of the results in the in-vitro experiment, describing a new in-vivo experiment with domain experts and a new language. The latter experiment is crucial to show that a coverage-oriented matching function like TR-ESA can be incorporated in an interactive tool to support users in solving a real-world data linking problem. Finally, this experiment provided additional evidence to support the hypothesis that TR-ESA can be easily adapted to handle new languages.

6.1. Cross-lingual Information Retrieval

From the point of view of information retrieval, in this paper we presented a method for the cross-lingual retrieval of a list of relevant documents in a language L, given a query formulated in a language L' ($L' \neq L$), with query and documents represented by short textual content.

There is an abundance of literature on cross-lingual Information Retrieval (CLIR). Two main categories of approaches can be identified: text-based [19, 40, 53] and concept-based [45, 25, 50] approaches. In text-based approaches, documents are retrieved (or compared) using a common word space after that a preliminary translation step is applied to queries or documents to

bridge the language gap. Thus, we refer to these approaches as *translation-based* approaches. In concept-based approaches, documents are retrieved (or compared) using a common concept space to represent their content. The retrieval model presented in our paper, i.e, TR-ESA, can be defined as a hybrid approach because it uses a concept space for representing the documents but after applying machine translation to bridge the language gap.

McCarley [29] demonstrated that retrieval effectiveness is more influenced by the translation direction (e.g., Italian-to-English, English-to-Italian) than the translation element (queries vs. documents). This result demonstrates that a crucial role is played by the translation process. In the meanwhile, machine translation algorithms have dramatically improved their performance. Statistical Machine Translation algorithms demonstrated to be more effective at large scale than other approaches (e.g., rule-based) [41] and are now adopted in the most popular machine translation tools (e.g., Google Translate[15], Microsoft Translator[16]). Some approaches have also proposed to use a pivot language [47] for obtaining a common document representation. In that case, a direct translation from the source language to the target one is not performed, but the two languages are represented in a third common language (e.g., English). Translation-based approaches have the advantage that they can handle the large number of languages covered by machine translation with limited adaptation effort (although at varying levels of quality depending on the languages).

Concept-based approaches adopt implicit or explicit concept models. The most prominent implementations of implicit concept models are Latent Semantic Indexing (LSI) [6] and Latent Dirichlet Allocation (LDA) [3]. Both LSI and LDA perform a dimensionality reduction of the word space, where the reduced dimensions are the implicit concepts used for indexing new documents. LSI and LDA can be exploited to face cross-lingual retrieval tasks by representing documents lexicalized in different languages in a shared concept space [28]. On the other side, some approaches that use explicit concept models in CLIR use a cross-lingual adaptation of the Explicit Semantic Analysis (ESA) model, named Cross-Language ESA (CL-ESA) by its creators [50][17]. As discussed in details in Section 2.2, in CL-ESA documents are represented using a concept-based vector space. As in TR-ESA, i.e., the ESA model used

[15]https://translate.google.com/

[16]http://www.microsoft.com/translator/

[17]A model equivalent to CL-ESA was also introduced by another approach to CLIR [45]. However, in this paper we use the more systematic presentation of CL-ESA found in [50]

in our approach, text fragments (queries and documents) are represented using semantic concepts defined by Wikipedia articles. The idea proposed by CL-ESA is to represent a text fragment in terms of Wikipedia concepts (as the ESA model already does) in its native language and then to switch from one language to another by using cross-lingual links between Wikipedia articles.

In our work, we introduced a TRanslation-based version of ESA (TR-ESA) that combines translation-based and concept-based approaches by using ESA to extract Wikipedia-based representations of textual descriptions. Thus, differently from other translation-based approaches that use word vectors, our method uses a concept-based vector space for document representation. In our experiments we have shown that, when documents are very short, TR-ESA outperforms approaches that use plain word vectors. On the contrary, differently from CL-ESA, we first translate descriptions in different languages in a pivot language (i.e., English), and then we apply ESA. Our method has the advantage that can handle textual descriptions in many languages much easier. In fact, we need to conduct the effort-intensive task of processing the Wikipedia corpus only for one language. Instead, CL-ESA requires that the Wikipedia corpus in every language that has to be semantically analyzed is processed. Furthermore, our approach overcomes the problem related to the heterogeneity of different versions of Wikipedia. Indeed, the experimental evaluation on CL-ESA by Sorg et al. was carried out on four different languages (i.e., English, German, French and Spanish) whose Wikipedia versions are quite accurate (i.e., English $\sim 5,000,000$ pages, German, French, Spanish more than $1,000,000$ pages). Conversely, in our work, we handled also languages with limited coverage in Wikipedia. For example, the Norwegian Wikipedia counts only $\sim 400,000$ pages[18]. An English Wikipedia concept that is not linked to any concept in the Norwegian Wikipedia represents only noise in document representations. The idea of introducing, in TR-ESA, a machine translation step before the ESA feature generation process revealed to be useful to extend the ESA benefits also to languages that have little coverage in Wikipedia (despite potential errors in the translation process).

6.2. Cross-lingual Data Linking

Data linking, frequently called also link discovery, is a task devoted to establish semantic links between resources in Knowledge Bases (KBs). Since

[18]https://en.wikipedia.org/wiki/List_of_Wikipedias

resources in a KB are represented as ontology instances and automatic approaches to data linking focus on *sameAs* links, i.e., links that tell us that two resources denote the same real world object, data linking can be assimilated to the task of *instance matching* in the ontology matching literature[19].

The problem of supporting data linking among resources lexicalized in different languages has received little interest[20]. In the NTCIR cross-lingual link discovery task, hyperlinks (not semantic links) have to be found between text fragments in documents and other relevant documents. However, it is worth mentioning that the CL-ESA model has been used for this task in a work that we have cited in the subsection about cross-lingual document retrieval [25]. Only two other approaches to this problem have been proposed so far [26, 27]. In both approaches, experimented in interlinking RDF data in English and Chinese, each resource is represented by a virtual document built by collecting every literal associated with this resource by a property path of maximum length equal to two. In [2] a system that uses semantic data for cross-lingual linking of news article clusters is proposed. The approach is mainly based on the identification of named entities into the news. The experimental results demonstrate that taking into account the semantic aspect of news increases performance and improves linking. In the news domain there is another work that compares different cross-lingual document similarity measures based on Wikipedia to establish link connections of articles in different languages [46]. Also this work is mainly based on the identification of named entities in the articles by annotating the entity with the corresponding Wikipedia page. However, in our experimental evaluations we demonstrated that those approaches are not sufficiently effective with short textual descriptions. Finally, in [54], the authors investigate the problem of linking information between different usages of the same languages, e.g. colloquial and formal idioms or the language of consumers versus the language of sellers. Hence, in this work the cross-lingual linking problem is specifically a cross-idiomatic linking task. The adopted approach is similar to that proposed in this paper: an information retrieval framework where a Web element si considered a query and other Web elements, from the target environment, are documents ranked according to the relevance query. This work compares three different LDA-based models. However,

[19]We remark that in every experiment discussed in this paper resources are represented as ontology instances of class *Service*.

[20]This problem should not be confused with the *cross-lingual link discovery* task addressed in the NTCIR conference [51]

since LDA models are statistical models that reduce the dimensionality of the document representation, they require large corpora.

The main contribution of our work to the emerging field of cross-lingual data linking is twofold. On the one hand, we present a concept-based cross-lingual matching function for short textual descriptions that can be incorporated in link specifications and that can be easily applied to a very large number of languages with minor configuration effort. We experimented our matching function with six languages other than English. The matching function of TR-ESA is a bounded similarity measure, which makes it easy to combine it with other measures. e.g., in linear weighted combinations [39]. While TR-ESA uses a concept-based vector space, methods proposed in previous work use a word-based vector space. When rich virtual documents can be collected for the resources, the word-based vector space model may be sufficient to compare different entities. Conversely, when only short textual descriptions are available for comparison, our experiments show that the feature generation process enabled by TR-ESA yields to better performance than the comparison of plain word-based representations. On the other hand, we propose a tool-supported methodology for data linking, which can be applied when only little evidence can be used for link discovery (only short textual descriptions, in our case), thus making it difficult to automatically decide whether linking or not. Our experiments with users show that, despite the difficulty of the task, the recommendation of potential links is effective to support users in linking their data. In addition, these experiments help us find out that relations other than *sameAs* can be considered when linking ontology instances. Instances can represent objects at different granularity levels [24], which convinced us to consider *skos:broadMatch* and *skos:narrowMatch* links. As a matter of fact, the problem of misused *sameAs* links in the semantic Web has been discussed also in previous work [17].

6.3. Cross-lingual Ontology Matching

Several other ontology matching problems have been investigated so far, including cross-lingual ontology matching, in addition to instance matching. Sometimes it may be difficult to decide what should be represented as an ontology concept or as an ontology instances on purely theoretical basis, and this choice is motivated by application requirements or data management concerns. However, ontology concepts are expected to be organized in hierarchies or subconcept graphs. Remarkably, the graph-based structure of an ontology is one of the key assets considered for matching ontology concepts or properties [48], i.e., schema-level ontology matching (in the following, we use *ontology matching* to refer solely to schema-level ontology matching).

A benchmark to evaluate cross-lingual ontology matching has been proposed in the context of the Ontology Alignment Evaluation Initiative (OAEI), with the aim of comparing ontology matching systems on shared and precisely defined test cases. The test cases used to evaluate cross-lingual ontology matching (Multifarm dataset[21]) are based on a subset of the Conference dataset (a collection of ontologies describing the *domain of conference organization*) translated into eight different languages (Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish) and the corresponding alignments between these ontologies. During the OAEI 2014 edition [10] only three systems [11, 23, 9] out of nine implemented a specific cross-lingual approach combined with structural-based matching strategies for the cross-lingual test case. All systems implementing cross-lingual techniques were based on machine translation tools and outperformed the other language-agnostic methods. Therefore, also for this task, the translation process demonstrated to be a key solution to improve the accuracy of the proposed systems. A recent paper has further investigated the effectiveness of machine translations for cross-lingual ontology matching at large scale, using Google Translate and BabelNet as translation resources [21]. The work presented in this paper cannot be considered a full-fledged cross-lingual ontology matching system because we consider only lexical matching between textual descriptions. If we cannot compare the performance of our matching method to the performance of these systems on the Multifarm dataset, we can discuss potential advantages and disadvantages of TR-ESA as a lexical matching method in the field of cross-lingual ontology matching. CIDER-CL [15] is a cross-lingual ontology matching system that adopts CL-ESA as a lexical matching method (for a detailed comparison between CL-ESA and TR-ESA we refer to Section 6.1). CIDER-CL obtains good results, but only on a limited number of matching tasks. In particular, the system could not run the matching tasks that use languages not covered by CL-ESA, thus suggesting that applying CL-ESA to a large number of languages requires a significant amount of effort. Two approaches search in the web [43] or in Wikipedia [22] for documents related to a concept label. The results are merged in a virtual document for comparison in a word-based vector space. In the first approach queries to the search engine are translated into a pivot language [43]. In the second approach, links between articles in localized Wikipedia versions are used to bridge the language gap [22] (similarly as in CL-ESA). Finally, a purely lexical method has been recently proposed for the

[21] http://www.irit.fr/recherches/MELODI/multifarm/

cross-lingual matching of lexical ontologies [20]; however, this method leverages the simultaneous translation of different synonyms associated with the ontology concepts to be matched, which are not available in several ontology matching tasks.

6.4. Cross-lingual Entity Linking

The problem of cross-lingual entity linking (EL) is related to, but also quite different from, the problem addressed in our work. Both problems require cross-lingual matching methods, and our work focuses on textual information. However, in EL the text fragment that has to be linked to a named entity is surrounded by additional text, which is used as context to support disambiguation in the linking task [33]. In our case, each textual description that has to be linked to a resource in a KB is the description of the resource itself. In other terms, we do not have additional text surrounding the portion of text that we use in the matching process.

However, in our experiments we also used state-of-the-art tools to link named entities to Wikipedia concepts, i.e. WikiMiner [32] and Tagme [13], or DBpedia entities, i.e., DBpedia Spotlight [30]. Since these tools are monolanguage or cover only few languages, we run these tools on textual descriptions in English returned by a machine translation tool. We show in our experiments that these methods are outperformed by our method, which makes use of TR-ESA to provide enriched representations of the resources.

Some native cross-lingual EL approaches have also been proposed. An approach analyzes and refines cross-lingual entity linking by using two methods, one based on the co-occurrence of entity mentions and one based on topic coherence [4]. This approach makes use of significant contextual information extracted from text to improve linking. Furthermore, the approach is supervised, while our TR-ESA-based method is unsupervised.

Babelfy is a method to jointly perform word sense disambiguation and EL [33]. The method is amenable to cross-lingual EL because it uses the BabelNet lexicon [38], which provides resource descriptions lexicalized in several languages. Since Babelfy uses the relations available in BabelNet as well as the coherence among several mentions in a text fragment to improve the EL task, we cannot apply the matching techniques used in Babelfy in our structureless matching scenario. Thus, the only option would be to use Babelfy to enrich the resource descriptions as we did in our experiments with other EL tools. We used Babelfy in our experiments and the difficulty of extracting entities from short descriptions look like a structural problem of any EL tool that we used in our experiments.

7. Conclusions and Future Work

In this paper we presented a cross-lingual link discovery approach based on an effective method to match short textual descriptions written in different languages. Our matching method is based on the definition of TR-ESA, a translation-based version of the Explicit Semantic Analysis that performs a machine translation of the input text and generates a Wikipedia-based representations for it. This matching method is used to recommend potential cross-lingual links to users of a web application by reducing their effort in completing difficult linking tasks, which require human intervention. To evaluate the approach we implemented a link discovery IR system named CroSeR and we carried out two experiments: an *in-vitro* experiment on five different languages and datasets, with several alternative approaches, and an *in-vivo* experiment on a sixth anguage involving 15 domain experts. Results of the two experiments are coherent and show the effectiveness of our approach in terms of coverage, accuracy, and ranking. Even though some of the languages used in the experiments have words in common (e.g. German and Dutch), the service descriptions from a language to another vary consistently. Furthermore, we investigated the performance also with languages very different from English, such as Italian[22]. Also in this case, the system achieves very good performance. This observation makes our results fairly generalizable. Indeed, our approach can be straightforwardly extended to any other language (even with poor Wikipedia version) by plugging the respective machine translation service in, although the accuracy of the recommendations may be affected by the quality of the translations. Furthermore, our matching method can be applied to any domain covered by the Wikipedia encyclopedic corpus, and it can be used as a recall-oriented cross-lingual matching method in contexts where only poor textual descriptions are available. It could be also combined with other matching methods, as it is often the case for individual matching functions.

As a future work we plan to analyze TR-ESA when different translation services are exploited, in order to determine whether differences performance are due to translation quality or to properties of specific languages. We also plan to develop another service based on the matching function described in this paper and currently implemented at a very preliminary stage, which retrieves candidate resources upon a query formulated in the native natural language of the user.

[22]Italian belongs to the Romance language family, whereas English, Dutch, German belong to the Germanic language family.

In relation to our cross-lingual interactive data linking approach, we want to incorporate methods that leverage links incrementally established by users to optimize the matching function in a pay-as-you-go fashion. We plan to adapt approaches proposed in the ontology matching and entity linking fields, which collect the feedback from individual or groups of users [7, 5].

To improve cross-lingual data linking for service descriptions, we plan to investigate methods to enrich the descriptions available in both source and target languages. In particular, we plan to implement an acronym expansion method that can be useful to resolve semantic mismatches that are not addressed by our matching method.

References

[1] Baldassarre, C., Cremaschi, M., and Palmonari, M. (2013). Bridging the gap between citizens and local administrations with knowledge-based service bundle recommendations. In *24th International Workshop on Database and Expert Systems Applications, DEXA 2013, Prague, Czech Republic, August 26-29, 2013*, pages 157–161.

[2] Belyaeva, E., Košmerlj, A., Muhič, A., Rupnik, J., and Fuart, F. (2015). Using semantic data to improve cross-lingual linking of article clusters. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:64–70.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022.

[4] Cassidy, T., Ji, H., Deng, H., Zheng, J., and Han, J. (2012). Analysis and refinement of cross-lingual entity linking. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 1–12. Springer.

[5] Cruz, I. F., Palmonari, M., Loprete, F., Stroe, C., and Taheri, A. (2015). Quality-based model for effective and robust multi-user pay-as-you-go ontology matching. *Semantic Web*, 7(4):463–479.

[6] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

[7] Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. (2013). Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *VLDB J.*, 22(5):665–687.

[8] Deshpande, M. and Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177.

[9] Djeddi, W. E. and Khadir, M. T. (2014). XMap++: Results for OAEI 2014. In [49], pages 163–169.

[10] Dragisic, Z., Eckert, K., Euzenat, J., Faria, D., Ferrara, A., Granada, R., Ivanova, V., Jiménez-Ruiz, E., Kempf, A. O., Lambrix, P., Montanelli, S., Paulheim, H., Ritze, D., Shvaiko, P., Solimando, A., dos Santos, C. T., Zamazal, O., and Grau, B. C. (2014). Results of the Ontology Alignment Evaluation Initiative 2014. In [49], pages 61–104.

[11] Faria, D., Martins, C., Nanavaty, A., Taheri, A., Pesquita, C., Santos, E., Cruz, I. F., and Couto, F. M. (2014). Agreementmakerlight results for OAEI 2014. In [49], pages 105–112.

[12] Fernando, S., Hall, M. M., Agirre, E., Soroa, A., Clough, P. D., and Stevenson, M. (2012). Comparing taxonomies for organising collections of documents. In Kay, M. and Boitet, C., editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proc. of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 879–894. Indian Institute of Technology Bombay.

[13] Ferragina, P. and Scaiella, U. (2010). TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia entities). In *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

[14] Gabrilovich, E. and Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research (JAIR)*, 34:443–498.

[15] Gracia, J. and Asooja, K. (2013). Monolingual and cross-lingual ontology matching with CIDER-CL: evaluation report for OAEI 2013. In Shvaiko, P., Euzenat, J., Srinivas, K., Mao, M., and Jiménez-Ruiz, E., editors, *Proc. of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013.*, volume 1111 of *CEUR Workshop Proc.*, pages 109–116. CEUR-WS.org.

[16] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., and McCrae, J. (2012). Challenges for the multilingual web of

data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71.

[17] Halpin, H., Hayes, P. J., McCusker, J. P., McGuinness, D. L., and Thompson, H. S. (2010). When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data. In *The Semantic Web–ISWC 2010*, pages 305–320. Springer.

[18] Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

[19] Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., and Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002. *Information Retrieval*, 7(1-2):99–119.

[20] Helou, M. A. and Palmonari, M. (2015). Cross-lingual lexical matching with word translation and local similarity optimization. In *Proc. of the 10th International Conference on Semantic Systems, SEMANTiCS 2015, Vienna, Austria, September*.

[21] Helou, M. A., Palmonari, M., and Jarrar, M. (2016). Effectiveness of automatic translations for cross-lingual ontology mapping. *Journal of Artificial Intelligence Research*, 55:165–208.

[22] Hertling, S. and Paulheim, H. (2012). WikiMatch - Using Wikipedia for Ontology Matching. In *Proc. of the 7th International Workshop on Ontology Matching (OM 2012)*.

[23] Jiménez-Ruiz, E., Grau, B. C., Xia, W., Solimando, A., Chen, X., Cross, V., Gong, Y., Zhang, S., and Chennai-Thiagarajan, A. (2014). LogMap family results for OAEI 2014? In *Proc. of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October*, volume 20, pages 126–134.

[24] Keet, C. M. (2009). From granulation hierarchy to granular perspective. In *The 2009 IEEE International Conference on Granular Computing, GrC 2009, Lushan Mountain, Nanchang, China, 17-19 August 2009*, pages 306–311. IEEE.

[25] Knoth, P., Zilka, L., and Zdrahal, Z. (2011). Using explicit semantic analysis for cross-lingual link discovery. In *5th International Workshop*

on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies, pages 687–696.

[26] Lesnikova, T., David, J., and Euzenat, J. (2014). Interlinking English and Chinese RDF Data Sets Using Machine Translation. In Völker, J., Paulheim, H., Lehmann, J., Sack, H., and Svátek, V., editors, *Proc. of the 3rd Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data co-located with 11th Extended Semantic Web Conference (ESWC 2014), Crete, Greece, May 25, 2014.*, volume 1243 of *CEUR Workshop Proc.* CEUR-WS.org.

[27] Lesnikova, T., David, J., and Euzenat, J. (2015). Interlinking English and Chinese RDF Data Using BabelNet. In *Proc. of the 2015 ACM Symposium on Document Engineering*, DocEng '15, pages 39–42, New York, NY, USA. ACM.

[28] Littman, M. L., Dumais, S. T., and Landauer, T. K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In *Cross-language information retrieval*, pages 51–62. Springer.

[29] McCarley, J. S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 208–214, Stroudsburg, PA, USA. Association for Computational Linguistics.

[30] Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia Spotlight: shedding light on the web of documents. In *Proc. of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA. ACM.

[31] Miles, A. and Pérez-Agüera, J. R. (2007). SKOS: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3-4):69–83.

[32] Milne, D. and Witten, I. H. (2008). Learning to link with Wikipedia. In *Proc. of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

[33] Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

[34] Narducci, F., Basile, P., Musto, C., Lops, P., Caputo, A., de Gemmis, M., Iaquinta, L., and Semeraro, G. (2016). Concept-based item representations for a cross-lingual content-based recommendation process. *Information Sciences*, 374:15–31.

[35] Narducci, F., Palmonari, M., and Semeraro, G. Cross-language semantic matching for discovering links to e-gov services in the lod cloud. *Proc. of the 2nd Int. Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD), co-located with the 10th (ESWC 2013), Montpellier, France, May 26-30, 2013.*

[36] Narducci, F., Palmonari, M., and Semeraro, G. (2013). Cross-language semantic retrieval and linking of e-gov services. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proc., Part II)*, pages 130–145.

[37] Navigli, R. and Ponzetto, S. P. (2012a). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

[38] Navigli, R. and Ponzetto, S. P. (2012b). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artif. Intell.*, 193:217–250.

[39] Ngonga Ngomo, A.-C. (2012). On link discovery using a hybrid approach. *Journal on Data Semantics*, 1(4):203–217.

[40] Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81. ACM.

[41] Och, F. J. (2005). Statistical machine translation: Foundations and recent advances. In *Tutorial at Tenth Machine Translation Summit.*

[42] Palmonari, M., Viscusi, G., and Batini, C. (2008). A semantic repository approach to improve the government to business relationship. *Data Knowl. Eng.*, 65(3):485–511.

[43] Paulheim, H. (2012). WeSeE-Match results for OEAI 2012. In *Proc. of the 7th International Workshop on Ontology Matching (OM 2012).*

[44] Peters, C., Braschler, M., and Clough, P. (2012). *Multilingual information retrieval: from research to practice*. Springer.

[45] Potthast, M., Stein, B., and Anderka, M. (2008). A Wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530. Springer.

[46] Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., and Grobelnik, M. (2016). News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, 55:283–316.

[47] Savoy, J. and Dolamic, L. (2009). How effective is Google's translation service in search? *Communications of the ACM*, 52(10):139–143.

[48] Shvaiko, P. and Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176.

[49] Shvaiko, P., Euzenat, J., Mao, M., Jiménez-Ruiz, E., Li, J., and Ngonga, A., editors (2014). *Proc. of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014*, volume 1317 of *CEUR Workshop Proc.* CEUR-WS.org.

[50] Sorg, P. and Cimiano, P. (2012). Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45.

[51] Tang, L.-X., Kang, I.-S., Kimura, F., Lee, Y.-H., Trotman, A., Geva, S., and Xu, Y. (2013). Overview of the NTCIR-10 cross-lingual link discovery task. In *Proc. of the Tenth NTCIR Workshop Meeting, page to appear, NII, Tokyo*.

[52] Voorhees, E. M. (1999). TREC-8 question answering track report. In *Proc. of the 8th Text Retrieval Conference*, pages 77–82.

[53] Yu, K. and Tsujii, J. (2009). Bilingual dictionary extraction from Wikipedia. In *Proc. of Machine Translation Summit XII*, pages 379–386.

[54] Zoghbi, S., Vulić, I., and Moens, M.-F. (2016). Latent dirichlet allocation for linking user-generated content and e-commerce data. *Information Sciences*, 367(C):573–599.