Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Tatiana Lesnikova

Thèse dirigée par **Jérôme Euzenat**
et codirigée par **Jérôme David**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Liage de données RDF: Evaluation d'approches interlingues

Thèse soutenue publiquement le **4 mai 2016**,
devant le jury composé de :

Prof. Laurent Besacier
Université Grenoble Alpes, France, Président
Prof. Aldo Gangemi
Université Paris 13, France, Rapporteur
Dr. Nathalie Aussenac-Gilles
MELODI-CNRS, France, Rapporteur
Dr. Jorge Gracia del Río
Universidad Politécnica de Madrid, Espagne, Examinateur
Dr. Jérôme Euzenat
INRIA, France, Directeur de thèse
Dr. Jérôme David
Université Grenoble Alpes, France, Co-Directeur de thèse

RDF Data Interlinking: Evaluation of Cross-lingual Methods

Abstract

The Semantic Web extends the Web by publishing structured and interlinked data using RDF. An RDF data set is a graph where resources are nodes labelled in natural languages. One of the key challenges of linked data is to be able to discover links across RDF data sets. Given two data sets, equivalent resources should be identified and linked by owl:sameAs links. This problem is particularly difficult when resources are described in different natural languages.

This thesis investigates the effectiveness of linguistic resources for interlinking RDF data sets. For this purpose, we introduce a general framework in which each RDF resource is represented as a virtual document containing text information of neighboring nodes. The context of a resource are the labels of the neighboring nodes. Once virtual documents are created, they are projected in the same space in order to be compared. This can be achieved by using machine translation or multilingual lexical resources. Once documents are in the same space, similarity measures to find identical resources are applied. Similarity between elements of this space is taken for similarity between RDF resources.

We performed evaluation of cross-lingual techniques within the proposed framework. We experimentally evaluate different methods for linking RDF data. In particular, two strategies are explored: applying machine translation or using references to multilingual resources. Overall, evaluation shows the effectiveness of cross-lingual string-based approaches for linking RDF resources expressed in different languages. The methods have been evaluated on resources in English, Chinese, French and German. The best performance (over 0.90 F-measure) was obtained by the machine translation approach. This shows that the similarity-based method can be successfully applied on RDF resources independently of their type (named entities or thesauri concepts). The best experimental results involving just a pair of languages demonstrated the usefulness of such techniques for interlinking RDF resources cross-lingually.

Liage de données RDF: Evaluation d'approches interlingues

Résumé

Le Web des données étend le Web en publiant des données structurées et liées en RDF. Un jeu de données RDF est un graphe orienté où les ressources peuvent être des sommets étiquetées dans des langues naturelles. Un des principaux défis est de découvrir les liens entre jeux de données RDF. Étant donnés deux jeux de données, cela consiste à trouver les ressources équivalentes et les lier avec des liens owl:sameAs. Ce problème est particulièrement difficile lorsque les ressources sont décrites dans différentes langues naturelles.

Cette thèse étudie l'efficacité des ressources linguistiques pour le liage des données exprimées dans différentes langues. Chaque ressource RDF est représentée comme un document virtuel contenant les informations textuelles des sommets voisins. Les étiquettes des sommets voisins constituent le contexte d'une ressource. Une fois que les documents sont créés, ils sont projetés dans un même espace afin d'être comparés. Ceci peut être réalisé à l'aide de la traduction automatique ou de ressources lexicales multilingues. Une fois que les documents sont dans le même espace, des mesures de similarité sont appliquées afin de trouver les ressources identiques. La similarité entre les documents est prise pour la similarité entre les ressources RDF.

Nous évaluons expérimentalement différentes méthodes pour lier les données RDF. En particulier, deux stratégies sont explorées: l'application de la traduction automatique et l'usage des banques de données terminologiques et lexicales multilingues. Dans l'ensemble, l'évaluation montre l'efficacité de ce type d'approches. Les méthodes ont été évaluées sur les ressources en anglais, chinois, français, et allemand. Les meilleurs résultats (F-mesure >0.90) ont été obtenus par la traduction automatique. L'évaluation montre que la méthode basée sur la similarité peut être appliquée avec succès sur les ressources RDF indépendamment de leur type (entités nommées ou concepts de dictionnaires).

Acknowledgements

I am genuinely grateful to my Ph.D. advisors Jérôme Euzenat and Jérôme David for their invaluable advice and encouragement from the first months and during the whole duration of my research. Their time, wise guidance, help, and patience allowed me to stay focused on the research topic and to overcome difficulties encountered on the way. Their supportive attitude inspired me to pursue research ideas and participate in scientific events. I am grateful to Jérôme David for his openness and help with programming. I am grateful to Jérôme Euzenat for a keen eye and critical reading of all my drafts and for seeing the writing of this manuscript through to completion. I very much enjoyed working in the EXMO team.

Within the Lindicle[1] project aimed at exploring multilingual data interlinking between French, English and Chinese data sources, I had a great chance to visit the Tsinghua University, Beijing, China in April 2015. I thank Juanzi Li and students from the Knowledge Engineering group for a warm welcome and for being generous hosts during my stay in Beijing.

I thank my friends for a great company while I was completing this thesis. My office mate Zhengjie introduced me to the life of a Ph.D. student when I first arrived in Grenoble. I thank Adam and Rosa together with Armen, Shreyas, Dmitry and Manuel for various interesting discussions and the good time that we shared. I thank Jean-Pierre, Paulette, Gilles, and Olga for support in stressful times and for making my life in France easier and more enjoyable. I also thank my friends from Saint-Petersburg for their emotional support and trust in me. I would like to thank my parents for creating conditions which stimulated me to embark on the Ph.D. project and to accomplish it. Their understanding and a sense of perspective help me to look forward.

Last but not the least, this work would not be possible without the doctoral grant from the University of Grenoble. I equally thank the National Research Agency (l'Agence Nationale de la Recherche) for funding the end of my Ph.D. through the Lindicle project. I also acknowledge assistance of the Inria's administrative staff, and, in particular, Marion Ponsot for her kind assistance in my numerous administrative matters.

[1] http://lindicle.inrialpes.fr/, ANR-12-IS02-0002

Contents

Chapter 1

Introduction

The purpose of computing is insight, not numbers.

— R.Hamming, Numerical Methods for Scientists and Engineers, 1962.

The development of communication technologies facilitates the publication of a vast amount of information on the Web. Information sources of a broad variety are created independently and distributed across heterogeneous repositories. As a consequence, identical resources can be described differently. Moreover, Web resources can be described using different natural languages. As an example, the fans from all over the world can describe a musical band using their own set of attributes as well as their native language. As a result, one would end up with different representations expressed in different natural languages referring to the same referent (a particular band). Given that there are thousands of common entities which are represented differently, it is important to provide technologies for connecting these data. Interlinking of resources across heterogeneous data sources is an important task in the Semantic Web in order to enhance semantic interoperability. Semantic Web technologies [13] offer the possibility to publish structured descriptions of entities according to a standard data model and to describe them using reusable vocabularies (ontologies). By publishing and interlinking structured data available online, it will be possible to aggregate knowledge about entities: different perspectives on semantically related entities will be brought together. This, in turn, would provide a "global" view on entities of interest.

According to the Semantic Web principles, data are published to allow automated processing. The Linked Data initiative aims at publishing structured and interlinked data at web scale by using semantic web technologies. These technologies provide different languages for expressing data as graphs (RDF), describing its organization through ontologies (OWL) and querying it (SPARQL) [55].

The four principles of Linked Oped Data have been defined by Tim Berners-Lee[1]:

1. Use URIs to identify things;
2. Use dereferenceable URIs;
3. Provide useful information for dereferenceable URIs;
4. Include links to other datasets.

Linked Open Data is a freely available data set collection expressed in RDF [52, 104]. The Linked Open Data Cloud (LOD)[2] contains several billion triples and several million interlinks. The data come from a broad variety of domains such as government, life sciences, media, geographic, and social.

This thesis mostly contributes to the LOD 4th principle since we aim at establishing links between identical resources from different RDF data sets. An owl:sameAs statement is used to link two identical resources due to the non-unique naming assumption (each RDF publisher uses its own identifiers).

This thesis addresses the problem of cross-lingual RDF data interlinking. The goal of our work is to evaluate methods to identify and link semantically related resources across RDF data sets in different languages. Given two RDF data sets with literals in different natural languages, the output will be a set of triples of type <URI owl:sameAs URI′>. For now, we restrict ourselves to owl:sameAs[3] link [50] as it is a classical type of link that is usually established, and it is also important for tracking information about the same resource across different data sources.

Despite the development of the Semantic Web, Internet is likely to continue to accommodate a diversity of natural languages. Even though there are many resources in English, some other languages occupy a decent portion of the Web space as well, see language statistics[4] in Figure 1.1. At present, the number of languages[5] of RDF data sets amounts to 474. Thus, the necessity to tackle the language heterogeneity problem will persist.

DBpedia[6] is a knowledge base, providing an RDF representation of Wikipedia, in which multiple language labels are attached to the individual concepts. It has become the nucleus for the Web of Data. Though there are interlingual links between different language versions of Wikipedia, there are knowledge bases in other languages which are not interlinked. For example, XLore [136] is an RDF Chinese knowledge base which provides a semantic representation of national

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://lod-cloud.net/
[3] http://www.w3.org/TR/owl-ref/#sameAs-def
[4] http://www.internetworldstats.com/stats7.htm
[5] http://stats.lod2.eu/languages
[6] http://wiki.dbpedia.org

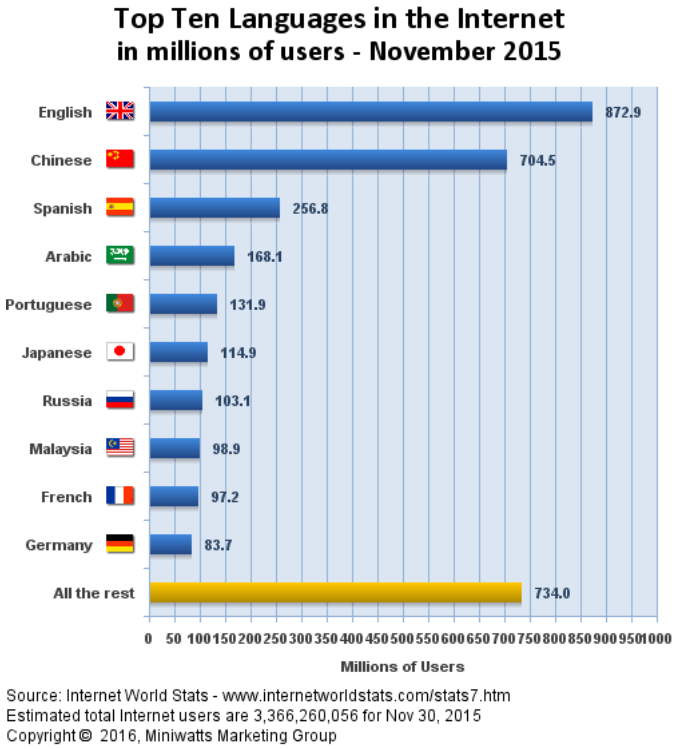


Figure 1.1: Internet world users by language statistics.

knowledge sources (Baidu baike, Hudong baike). Other publishers such as the French National Library [117], the Spanish National Library [130], the British Museum[7] make their data available using RDF model in their own language. Overall, there are many resources to be interlinked in the LOD cloud.

The growing number of RDF data sources with multilingual labels and the importance of cross-lingual links for other Semantic Web applications motivate our interest in cross-lingual link discovery.

One of the key challenges of linked data is to be able to discover links across datasets [34]. This problem is particularly difficult when entities are described in different natural languages on which string similarity measures cannot be applied directly. Hence, other approaches for bridging languages must be considered. The importance of cross-lingual data interlinking has been discussed in several works [16, 46, 47]. Recently a Best Practices for Multilingual Linked Open Data Community Group[8] has been created to elaborate a large spectrum of practices with regard to multilingual LOD.

Cross-lingual interlinking consists in discovering links between identical resources across diverse RDF sources in different languages, see Figure 2.5. It is

[7] http://collection.britishmuseum.org/

[8] http://www.w3.org/community/bpmlod/

particularly challenging due to several reasons:

(a) the structure of graphs can be different and the structure-based techniques may not be of much help;

(b) even if the structures are similar to one another, the properties themselves and their values are expressed in different natural languages.

The approaches proposed in this thesis deal with symbolic information extracted from RDF graphs: the values of properties are usually natural language words. We adopt a Natural Language Processing (NLP) approach to address the problem of finding the same object described in two different languages.

The contribution of this thesis is a study of techniques for cross-lingual data interlinking. To evaluate such techniques, a general framework for interlinking identical RDF resources is first proposed. This framework can be viewed as a tool for evaluating the techniques. The main features of the proposed framework are:

(a) an RDF data set can be described only in one natural language, no multilinguality is required;

(b) the approaches work without prior ontology matching;

(c) the framework includes several modifiable parameters which are tested during evaluation.

The obtained results depend on two components: (1) the resource representation containing symbolic information from graphs; (2) application of language-specific techniques to these representations. In particular, we investigate the impact of machine translation and multilingual lexicon mapping on the resource comparison. The best F-measure results using machine translation exceed 0.90 on a language pair as distinct as English-Chinese.

Availability of cross-lingual links is imperative for several neighboring research areas. For example, to overcome the problem of ontology heterogeneity, some research has been done on monolingual ontology integration based on instances interlinked by owl:sameAs [139]. If owl:sameAs links could be provided between instances expressed in different languages, other experiments on integrating underlying ontologies could be conducted. The owl:sameAs links between instances can be also valuable in applications such as Question Answering over multilingual structured knowledge-base [18] since a system can take advantage of the information presented in a language different from a language that is being queried. Hence, links between corresponding elements of the heterogeneous sources facilitate the integration of Web data and the uniform access to heterogeneous repositories.

The rest of the thesis is structured as follows. Chapter 2 focuses on the cross-lingual interlinking problem and research questions which are addressed in the present research. The chapter provides preliminaries on RDF graphs and cross-lingual graphs in particular. It clarifies what information can be used for their interlinking. Chapter 3 provides an overview of the state of the art and related research in neighboring areas. The problem of object matching has been studied in several fields such as databases, cross-lingual information retrieval, multilingual ontology matching. Advantages and disadvantages of approaches to tackle information in different languages are discussed. The present research is also classified according to existing classifications of matching techniques in the Semantic Web. Chapter 4 describes a general framework for interlinking RDF resources described in different languages. The framework incorporates a chain of processes that are parameterized. This framework allows to evaluate cross-lingual techniques in a systematic way. The experiments relying on machine translation are described in Chapters 5 and 7. In Chapter 5, a translation-based interlinking method is proposed and evaluated on the entities described in the English and Chinese languages. All entities represent named entities, e.g., actors, geographical places. However, we consider our method applicable to any type of Web resources. In order to verify that the performance of the approach does not depend on the presence of a name of a named entity, several experiments are conducted on a different type of data. Chapter 6 presents an interlinking method based on an external multilingual lexicon. This BabelNet-based method is compared to the machine translation method. Chapter 7 deals with thesauri matching. Several methods for interlinking general concepts from multilingual thesauri are evaluated. The concepts from the TheSoz thesaurus are described in three languages: English, French and German. The chapter also contains the evaluation of the translation-based method on concepts from EuroVoc in English and AGROVOC in Chinese. Chapter 8 describes perspectives on cross-lingual data interlinking. In particular, the design of an experiment for testing the hypothesis that the amount of textual information needed for resource interlinking depends on the nature of the resources. It also describes possible scenarios for combining machine translation and lexicon-based methods. Both methods could be complementary and may counterbalance each other. Finally, the conclusions are formulated in Chapter 9.

Chapter 2

Interlinking RDF Data in Different Languages

Abstract. In this chapter, we introduce the problem of cross-lingual data interlinking. Our research goal is to evaluate cross-lingual techniques which could facilitate linking different graphs with literals in different languages. We adopt a language-oriented approach and consider textual labels in graphs.

Qui se ressemble s'assemble.

— French proverb

The same knowledge can be expressed by different people, in different ways, and in different natural languages. This state of matters makes communication not very easy. However, even with the development of communication technologies, the same problems continue to hold their positions.

With the progress in global interconnectivity, communicating systems need to easily access to a variety of data sources in order to retrieve relevant information about resources. However, heterogeneity can be an obstacle to such access. There are several types of heterogeneity described in the literature [62, 97, 98, 115], in particular:

- Syntactic heterogeneity: differences in machine-readable aspects of representation and encodings of data;
- Structural heterogeneity: differences in metadata standards;
- Semantic heterogeneity: the same meaning of the data can be expressed in different ways.

These types of heterogeneity are present in the Semantic Web. At the syntactic level, heterogeneity is resolved by encoding knowledge in RDF and using Unicode. The use of various schemes and languages for describing RDF knowledge can lead to structural heterogeneity. In this thesis, we deal with the problem of semantic heterogeneity, i.e., the same knowledge can be described differently by different data providers, in particular, the descriptions can be provided in different natural languages.

Section 2.1 presents the preliminaries about the RDF model and shows how it is used to represent knowledge. Section 2.2 illustrates the problem of cross-lingual interlinking. We argue that interlinking can be based on language elements collected from knowledge descriptions. Research goals and questions which are answered throughout this thesis are specified in Sections 2.3 and 2.4. Sections 2.5 specifies the assumptions used in our study.

2.1 Resource Description Framework

Information is scattered on the Web. And this also holds for the Semantic Web. The Semantic Web provides technologies such as the Resource Description Framework (RDF) [65] for representing data on the web. Due to RDF, information on the Web can be turned from the unstructured collection into the structured data. RDF is a W3C data model according to which a resource is described by triples. A triple consists of a subject, a predicate, and an object. A predicate relates a subject to an object. Each subject and predicate (and optionally, object) component of an RDF statement is identified by a Uniform Resource Identifier (URI)[1] or a blank node[2]. An object can be also a literal: a Unicode string with optional language tags. Since data sets are created by publishers independently, there can be several URIs denoting the same resource across different RDF data sets. As a result, one needs to address the problem of entity resolution: identify and interlink the same entity across multiple data sources. An RDF resource description can remind feature structures used to represent linguistic knowledge as feature graphs [61] and conceptual graphs [122].

RDF statements form a directed labeled graph where the graph nodes represent resources and the edges represent typed relations between these resources. A set of statements about a resource constitutes a description set which contains certain characteristics of a resource and thus can ground the resource "identity". In our framework, we restrain the definition of RDF as a graph + identifiers (labels), see Figure 2.1. The identification of resources can be based on graph

[1] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#dfn-URI-reference
[2] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#dfn-blank-node

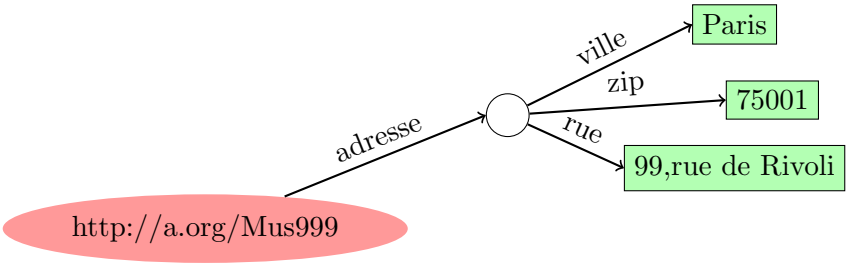


Figure 2.1: The directed labeled graph. The ellipse represents a resource identified by the URI, the round circle represents a blank node, and the boxes represent labels. Arrows represent relations.

structure and node labels. However, this problem can become particularly difficult when there are multilingual elements in a graph: string matching techniques can be inefficient. Hence, language-oriented techniques must be considered.

Definition of an RDF Graph: An RDF graph $\boldsymbol{G}$ is a set of triples (s, p, o) where s $\in \boldsymbol{U} \cup \boldsymbol{B}$; p $\in \boldsymbol{U}$ and o $\in \boldsymbol{U} \cup \boldsymbol{B} \cup \boldsymbol{L}$. Here $\boldsymbol{U}$ stands for URIs, $\boldsymbol{B}$ – for blank nodes and $\boldsymbol{L}$ – for literals (strings). A triple (s, p, o) forms a statement in which s is the subject, p is the predicate and o is the object of this triple.

In the context of RDF, two types of properties should be distinguished:

Datatype property: a predicate p is called a datatype property in $\boldsymbol{G}$ if in any triple (s, p, o) the object $o \in \boldsymbol{L}$.

Object property: a predicate p is called an object property in $\boldsymbol{G}$ if in any triple (s, p, o) the object $o \in \boldsymbol{U} \cup \boldsymbol{B}$.

RDF is usually expressed in a concrete serialization format. There are several formats[3] which allow to write RDF in a compact text form. A document containing data expressed in one of the formats is a textual representation of an RDF graph. To illustrate, Figure 2.2 shows an N-Triple document. It contains a sequence which represents the subject, predicate and object of an RDF triple. The sequence is terminated by a dot '.' A set of N-Triples can be converted[4] into an RDF/XML document as shown in Figure 2.3. Finally, it can be visualized[5] as a graph in a human-friendly form which is easier to read by humans as depicted in Figure 2.4.

In Figure 2.3, there are some properties with "rdfs" prefix, e.g., "rdfs:label", "rdfs:comment". These properties make part of RDF Schema[6].

[3]http://www.w3.org/TR/rdf-syntax-grammar/
[4]http://rdf-translator.appspot.com/
[5]https://www.w3.org/RDF/Validator/
[6]https://www.w3.org/TR/rdf-schema/

```
<http://dbpedia.org/resource/Mont_Blanc>    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>   <http://dbpedia.org/ontology/Mountain> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/elevation> "4810.0"^^<http://www.w3.org/2001/XMLSchema#double> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/firstAscentYear>
"1786-01-01T00:00:00+02:00"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/firstAscentPerson> <http://dbpedia.org/resource/Michel-Gabriel_Paccard> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/abstract>  "Mont Blanc (French pronunciation: \u200B[m\u0254\u0303.bl\u0251\u0303]) or
Monte Bianco (Italian pronunciation: [\u02C8monte \u02C8bja\u014Bko]), both meaning \"White Mountain\", is the highest mountain in the Alps and the European
Union. It rises 4,810 m (15,781 ft) above sea level and is ranked 11th in the world in topographic prominence. It is also sometimes known as La Dame blanche
(French for \"the White Lady\") or Il Bianco (Italian for \"the White One\").The mountain lies in a range called the Graian Alps, between the regions of Aosta
Valley, Italy, and Haute-Savoie, France. The location of the summit is on the watershed line between the valleys of Ferret and Veny in Italy and the valleys of
Montjoie Valley and Arve Valley in France. The Mont Blanc massif is popular for mountaineering, hiking, skiing, and snowboarding. Begun in 1957 and completed
in 1965, the 11.6 km (7\u00BC mi) Mont Blanc Tunnel runs beneath the mountain between these two countries and is one of the major trans-Alpine transport
routes." .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/mountainRange> <http://dbpedia.org/resource/Graian_Alps> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/locatedInArea> <http://dbpedia.org/resource/Italy> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://www.w3.org/2000/01/rdf-schema#label>    "Mont Blanc" .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/firstAscentPerson> <http://dbpedia.org/resource/Jacques_Balmat> .
<http://dbpedia.org/resource/Mont_Blanc>    <http://dbpedia.org/ontology/locatedInArea> <http://dbpedia.org/resource/France> .
<http://dbpedia.org/resource/France>    <http://dbpedia.org/ontology/leaderTitle>   "President" .
<http://dbpedia.org/resource/France>    <http://dbpedia.org/ontology/officialLanguage>  <http://dbpedia.org/resource/French_language> .
<http://dbpedia.org/resource/Jacques_Balmat>    <http://dbpedia.org/ontology/birthYear> "1762-01-01T00:00:00+02:00"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/Jacques_Balmat>    <http://dbpedia.org/ontology/abstract>  "Jacques Balmat, called le Mont Blanc (1762 \u2013 1834) was a
mountaineer, a Savoyard mountain guide, born in the Chamonix valley of the Savoy region, at this time part of the Duchy of Savoy.A chamois hunter and collector
of crystals, Balmat completed the first ascent of Mont Blanc with physician Michel-Gabriel Paccard on 8 August 1786. For this feat, the king Victor Amadeus III
gave him the honorary title le Mont Blanc.Balmat and Paccard's ascent of Mont Blanc was a major accomplishment in the early history of mountaineering. C.
Douglas Milner wrote \"The ascent itself was magnificent; an amazing feat of endurance and sustained courage, carried through by these two men only, unroped
and without ice axes, heavily burdened with scientific equipment and with long iron-pointed batons...\" .
<http://dbpedia.org/resource/Michel-Gabriel_Paccard>    <http://www.w3.org/2000/01/rdf-schema#comment>  "Michel Gabriel Paccard (1757\u20131827) was a Savoyard
doctor and alpinist, citizen of the Kingdom of Piedmont-Sardinia.Born in Chamonix, he studied medicine in Turin. Due to his passion for botany and minerals, he
met Horace-B\u00E9n\u00E9dict de Saussure, who initiated the race to be the first to ascend Mont Blanc.Gaston R\u00E9buffat wrote \"Like Saussure a devotee of
the natural sciences, he has a dream: to carry a barometer to the summit and take a reading there.\" .
<http://dbpedia.org/resource/France>    <http://dbpedia.org/ontology/longName>  "French Republic" .
<http://dbpedia.org/resource/France>    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>   <http://dbpedia.org/ontology/Country> .
<http://dbpedia.org/resource/Jacques_Balmat>    <http://dbpedia.org/ontology/deathYear> "1834-01-01T00:00:00+02:00"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/Michel-Gabriel_Paccard>    <http://dbpedia.org/ontology/deathYear>
"1827-01-01T00:00:00+02:00"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/France>    <http://dbpedia.org/ontology/anthem>    <http://dbpedia.org/resource/La_Marseillaise> .
<http://dbpedia.org/resource/France>    <http://dbpedia.org/ontology/capital>   <http://dbpedia.org/resource/Paris> .
<http://dbpedia.org/resource/Michel-Gabriel_Paccard>    <http://dbpedia.org/ontology/nationality>   <http://dbpedia.org/resource/Italy> .
<http://dbpedia.org/resource/Michel-Gabriel_Paccard>    <http://dbpedia.org/ontology/birthYear>
"1757-01-01T00:00:00+02:00"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/France>    <http://www.w3.org/2000/01/rdf-schema#label>    "France" .
<http://dbpedia.org/resource/Michel-Gabriel_Paccard>    <http://purl.org/dc/elements/1.1/description>   "Italian mountain climber" .
<http://dbpedia.org/resource/Jacques_Balmat>    <http://www.w3.org/2000/01/rdf-schema#label>    "Jacques Balmat" .
<http://dbpedia.org/resource/Michel-Gabriel_Paccard>    <http://www.w3.org/2000/01/rdf-schema#label>    "Michel-Gabriel Paccard" .
```

Figure 2.2: An example of an N-Triple document.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:dbo="http://dbpedia.org/ontology/"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
>
  <rdf:Description rdf:about="http://dbpedia.org/resource/France">
    <rdf:type rdf:resource="http://dbpedia.org/ontology/Country"/>
    <dbo:anthem rdf:resource="http://dbpedia.org/resource/La_Marseillaise"/>
    <dbo:leaderTitle>President</dbo:leaderTitle>
    <dbo:longName>French Republic</dbo:longName>
    <rdfs:label>France</rdfs:label>
    <dbo:capital rdf:resource="http://dbpedia.org/resource/Paris"/>
    <dbo:officialLanguage rdf:resource="http://dbpedia.org/resource/French_language"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Michel-Gabriel_Paccard">
    <dbo:nationality rdf:resource="http://dbpedia.org/resource/Italy"/>
    <rdfs:label>Michel-Gabriel Paccard</rdfs:label>
    <dc:description>Italian mountain climber</dc:description>
    <dbo:deathYear rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1827-01-01T00:00:00+02:00</dbo:deathYear>
    <rdfs:comment>Michel Gabriel Paccard (1757–1827) was a Savoyard doctor and alpinist, citizen of the Kingdom of
Piedmont-Sardinia.Born in Chamonix, he studied medicine in Turin. Due to his passion for botany and minerals, he met
Horace-Bénédict de Saussure, who initiated the race to be the first to ascend Mont Blanc.Gaston Rébuffat wrote "Like Saussure
a devotee of the natural sciences, he has a dream: to carry a barometer to the summit and take a reading there.</rdfs:comment>
    <dbo:birthYear rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1757-01-01T00:00:00+02:00</dbo:birthYear>
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Mont_Blanc">
    <dbo:locatedInArea rdf:resource="http://dbpedia.org/resource/Italy"/>
    <dbo:mountainRange rdf:resource="http://dbpedia.org/resource/Graian_Alps"/>
    <rdf:type rdf:resource="http://dbpedia.org/ontology/Mountain"/>
    <dbo:firstAscentYear rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1786-01-01T00:00:00+02:00</dbo:firstAscentYear>
    <dbo:abstract>Mont Blanc (French pronunciation:  [mɔ̃.blɑ̃]) or Monte Bianco (Italian pronunciation: ['monte 'bjaŋko]),
both meaning "White Mountain", is the highest mountain in the Alps and the European Union. It rises 4,810 m (15,781 ft) above
sea level and is ranked 11th in the world in topographic prominence. It is also sometimes known as La Dame blanche (French for
"the White Lady") or Il Bianco (Italian for "the White One").The mountain lies in a range called the Graian Alps, between the
regions of Aosta Valley, Italy, and Haute-Savoie, France. The location of the summit is on the watershed line between the
valleys of Ferret and Veny in Italy and the valleys of Montjoie Valley and Arve Valley in France. The Mont Blanc massif is
popular for mountaineering, hiking, skiing, and snowboarding. Begun in 1957 and completed in 1965, the 11.6 km (7¼ mi) Mont
Blanc Tunnel runs beneath the mountain between these two countries and is one of the major trans-Alpine transport
routes.</dbo:abstract>
    <dbo:firstAscentPerson rdf:resource="http://dbpedia.org/resource/Michel-Gabriel_Paccard"/>
    <dbo:locatedInArea rdf:resource="http://dbpedia.org/resource/France"/>
    <dbo:firstAscentPerson rdf:resource="http://dbpedia.org/resource/Jacques_Balmat"/>
    <rdfs:label>Mont Blanc</rdfs:label>
    <dbo:elevation rdf:datatype="http://www.w3.org/2001/XMLSchema#double">4810.0</dbo:elevation>
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Jacques_Balmat">
    <rdfs:label>Jacques Balmat</rdfs:label>
    <dbo:birthYear rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1762-01-01T00:00:00+02:00</dbo:birthYear>
    <dbo:abstract>Jacques Balmat, called le Mont Blanc (1762 – 1834) was a mountaineer, a Savoyard mountain guide, born in the
Chamonix valley of the Savoy region, at this time part of the Duchy of Savoy.A chamois hunter and collector of crystals,
Balmat completed the first ascent of Mont Blanc with physician Michel-Gabriel Paccard on 8 August 1786. For this feat, the
king Victor Amadeus III gave him the honorary title le Mont Blanc.Balmat and Paccard's ascent of Mont Blanc was a major
accomplishment in the early history of mountaineering. C. Douglas Milner wrote "The ascent itself was magnificent; an amazing
feat of endurance and sustained courage, carried through by these two men only, unroped and without ice axes, heavily burdened
with scientific equipment and with long iron-pointed batons...</dbo:abstract>
    <dbo:deathYear rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1834-01-01T00:00:00+02:00</dbo:deathYear>
  </rdf:Description>
</rdf:RDF>
```

Figure 2.3: An example of an RDF/XML document.

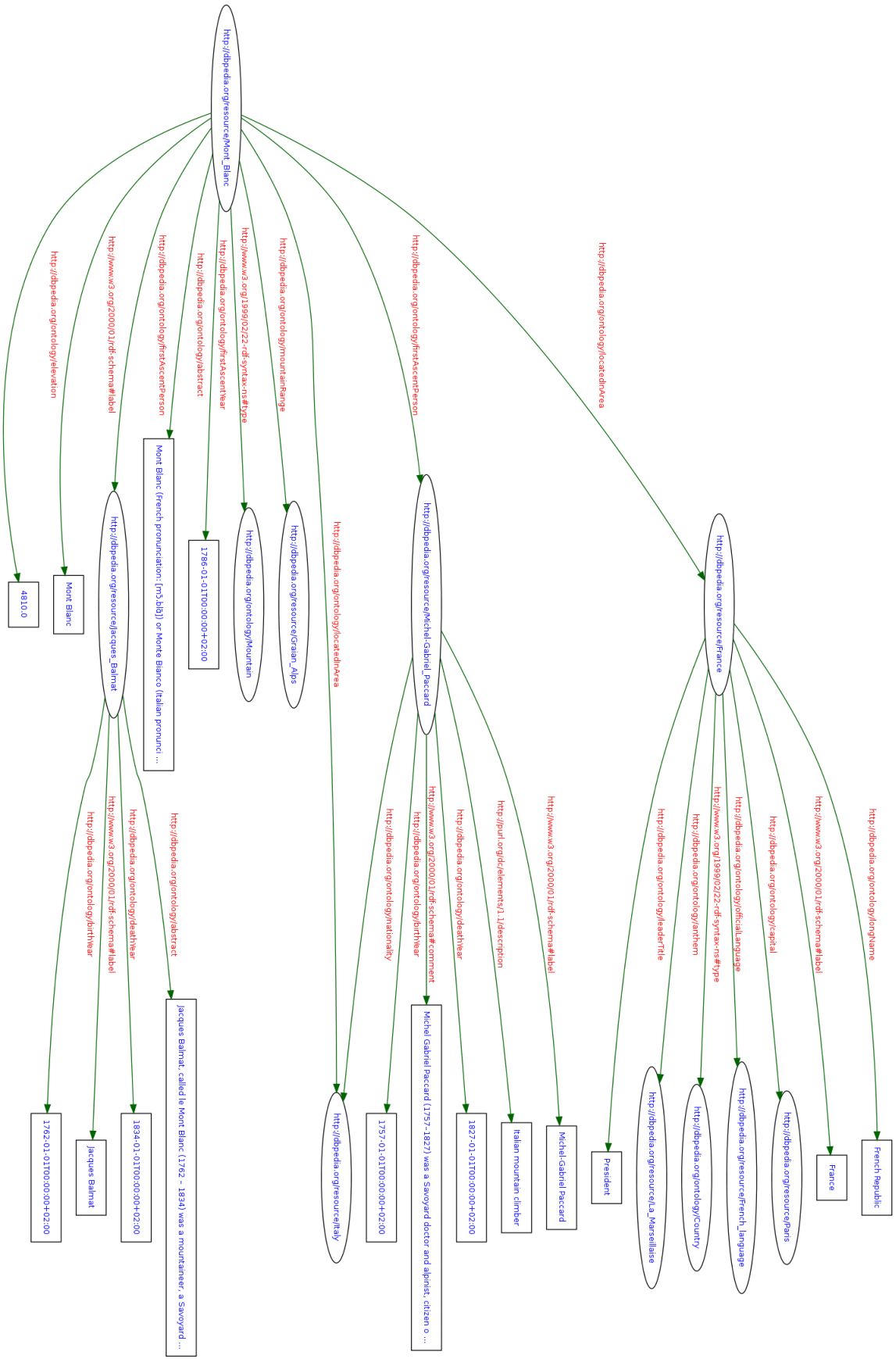


Figure 2.4: Visualization of an RDF graph. Ellipses contain URIs whereas boxes contain literals.

2.2 Cross-lingual RDF Data Interlinking

In Ontology Matching (OM), there exist a **distinction between multilingual matching and cross-lingual matching**. Three types of ontology matching are defined [123]:

1. Monolingual OM: ontology concepts are matched in a single language, e.g., the terms of both ontologies are in French;

2. Multilingual OM: ontology concepts are matched at least in two common languages;

3. Cross-lingual OM: ontology concepts are matched either by translating the source langue into the target one, or translating the target language into the source one; or translating both source and target languages into a pivot language. The number of languages to be translated can be more than one.

A slightly different differentiation between multilingual and cross-lingual ontology matching can be found in [32]. It makes multilingual matching more general by allowing to translate terms, i.e., ontology concepts are matched using multiple translated terms. However, cross-lingual matching is narrowed down to ontology matching with labels in two different languages.

We adopt this distinction in this thesis. This thesis addresses the RDF data interlinking cross-lingually. All conducted experiments include comparisons between two different languages.

Problem description. Given two RDF data sets with resources described in different natural languages, identify the same entities represented in different data sets and link them using owl:sameAs links. As a simple example, two graphs with multilingual elements to be interlinked are shown in Figure 2.5.

On the basis of this example, the following observations can be made concerning the elements on which similarity can rely:

- *URI's*: Two different URIs identify potentially identical resources, so it is not possible to rely on URIs. This is why data interlinking is necessary.
- *Structure*: The graph structures are different. However, if the structures were the same, properties and their values (literals) would be still in different languages.
- *Literals*: The datatype property values are literals in different languages. These linguistic elements can be used for comparing resources.
- *Ontology*: Resources can be described with the same ontology. Resources belonging to the same ontological class can be compared between themselves. This could reduce the number of possible comparisons as there is

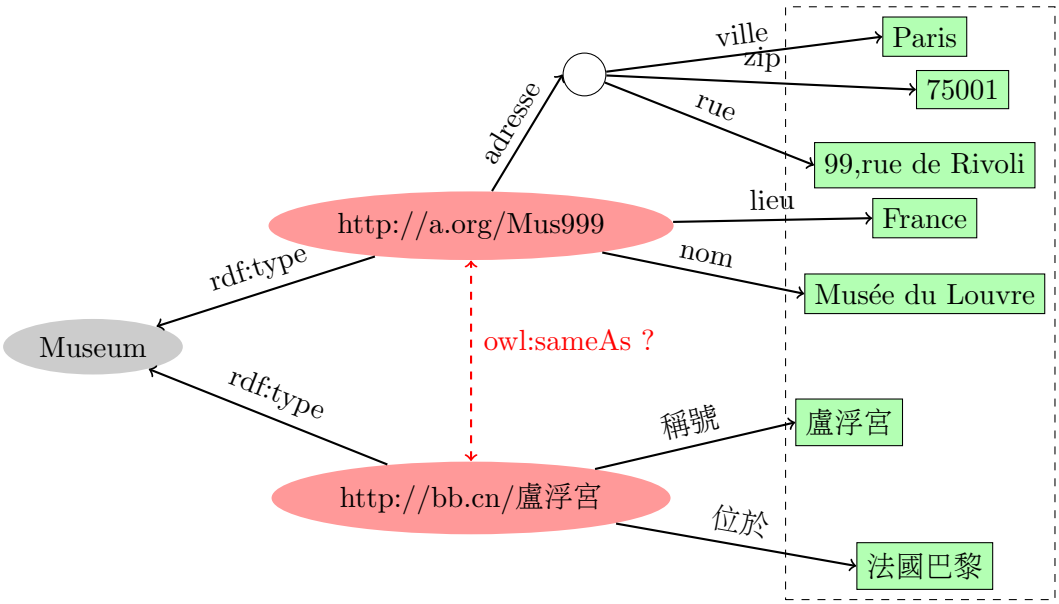


Figure 2.5: Interlinking RDF resources described in different natural languages. Two ellipses represent resources, an ellipse labelled "Museum" represents an ontological class which these resources belong to, the arrows represent predicates, and boxes represent objects with literals.

> no sense to compare totally different resources, e.g., museums and animals. In the example, both resources belong to the Museum class. Even though ontology provides information that both of these resources describe museums, it is still not enough to conclude that it is the same museum.

In this thesis, we design an interlinking framework which takes advantage of language elements in a graph. The proposed cross-lingual string based method relies on textual annotations associated with each resource, i.e., the comparison is based on literals.

2.3 Goals

Our research goal is to assess the suitability of NLP techniques for cross-lingual data interlinking.

We develop an approach in which RDF resources are represented as text documents and then compared. We apply standard Natural Language Processing (NLP) techniques (document preprocessing, term weights, similarity measures) on these documents. Considering that RDF resources are described in different languages, we particularly explore two strategies [66]:

- Applying Machine Translation (MT) in cross-lingual RDF data interlink-

ing [67];
- Using references to external multilingual resources [68].

To achieve this goal, we also pursue the following aims:

1. Identify data sets which can be used for experiments;
2. Build test sets (i.e., RDF test sets in different languages with a set of reference links);
3. Generate cross-lingual links;
4. Evaluate the performance of the proposed approach.

2.4 Research Questions

Our general *research question* is: To what extent is it possible to interlink data sets in different languages? To answer this question, within the framework described in Chapter 4, we need to explore which parameters influence this task. More specifically:

1. How to represent entities from RDF graphs?
    - How many language elements shall be collected from graphs?
2. How to make entities described in different natural languages comparable?
    - Is machine translation an appropriate medium to identify resource in two different languages?
    - Is a multilingual lexicon an appropriate medium to identify resource in two different languages?
    - What method performs better: a method based on translation technology or multilingual lexicon?
    - What is the impact of translating one language into another or pivot language?
    - How does the output of similarity measures vary according to the context?

All these parameters are studied with respect to specific contexts (language pairs, data set types, amount of textual data available).

2.5 Assumptions

Our work applies under the following assumptions about the techniques presented in this thesis:

1. The data to be linked are represented as a graph.
2. An RDF graph is labelled in natural languages. We assume the presence of language elements in a data set (properties and values). The methods are not suitable for RDF graphs containing purely numerical data.
3. The same natural language is used within one dataset.

2.6 Summary

Linking identical RDF resources is an interesting and a difficult problem to address. In the Semantic Web, knowledge is represented as a graph making the graph linking process significantly different from the traditional document comparison. Moreover, the same knowledge can be expressed in different languages which requires application of language-specific techniques in order to find correct correspondences from both languages. We introduced the cross-lingual interlinking problem which involves linking identical resources described in different natural languages from two RDF data sets. There are different elements in a graph on which the similarity between resources can be computed. This thesis investigates cross-lingual data interlinking based on literals. As literals are taken as a basis for resource comparison, this explains the choice of the NLP approach which is described in detail in Chapter 4.

The next chapter reviews state of the art and recent research efforts in cross-lingual data interlinking.

Chapter 3

State of the Art

Abstract. In this chapter, we review NLP techniques which allow to compare information in different languages. The problem of finding the same object across languages has been studied in many fields. Overcoming the language barrier may be simple or require external language resources.

Different domains study the problem of bringing together information about the same entity from multiple sources or searching for the same entity across multiple sources. We identified several of them. Each domain deals with resources represented as a database record or a raw text or a graph. In databases, this problem is known as record linkage. In Natural Language Processing (NLP), this problem is addressed in entity resolution, plagiarism detection, and cross-lingual information retrieval.

We singled out three main approaches to deal with information in different languages from these domains. Approaches relying on syntactic similarity between languages are reviewed in Section 3.2. Approaches which create an intermediate (interlingual) representation of textual content are described in Section 3.3. Finally, translation-based approaches are presented in Section 3.4. Multilingual resources which can be used to bridge the language barrier are assembled in Section 3.5.

In the Semantic Web, there are two domains which deal with object reconciliation. Notably, ontology matching aims at establishing correspondences between equivalent classes of different ontologies (see Section 3.6.1) and data interlinking where our research topic belongs to (see Section 3.6.2).

The backbone of our research is the multilingual aspect of representations of the same object. As such, methods and techniques used to bridge the gap across languages are emphasized.

3.1 Positioning with Respect to Other Fields

The problem of searching for the same entity across multiple sources dates back to the 1960s. In **databases**, the problem of finding information related to the same entity (person, place, etc.) from different sources is known under different names such as record linkage [33], deduplication [118], name matching [12], instance identification, record matching or the merge/purge problem [53]. For data integration purposes, information related to the same resource needs to be aggregated. The "duplicate record detection" is studied in [31] and a thorough survey is provided on the matching techniques. Many methods rely on character-based similarity, i.e., edit distance, but they are not appropriate for records in different languages. Though there has been much work done on record linkage, most of it concerns approaches for entities described in the same language. Very few research efforts have been dedicated to the problem of cross-lingual record linkage. Record linkage is related to our research in the sense that the duplicate RDF resources from heterogeneous data sources should be detected, whereas the search for duplicate records is done within a single data source complying to the same schema. Also, it contains neither the cross-lingual aspect nor RDF semantics or ontologies.

In NLP, the problems of entity resolution and cross-document coreference resolution [7] gained a close attention due to their complexity and importance for Information Retrieval, Question Answering, etc. The task of **entity resolution** is to find out whether the occurrences of a name in different natural language texts refer to the same object. There is no general solution to this problem, and the decision whether two names refer to the same entity usually relies on contextual clues. The research object of coreference resolution is a raw text, while in our case it is a graph, in which knowledge is split across this graph, i.e., knowledge is expressed in the form of graph structure and property values.

Cross-lingual entity linking has been addressed in the Knowledge Base Population track (KBP2011)[60] within the Text Analysis conference. The task of this track is to link entity mentions in a text to their counterparts in a knowledge base (Wikipedia). An entity mention is represented as a character string, so no RDF model is used for entity representation. If entity mentions are not in the knowledge base, they should be clustered into a separate group. Experiments were done both on monolingual (English) and cross-lingual (Chinese to English) data. Both language-independent and translation-based methods were used for that purpose [84].

In the field of Information Retrieval (IR), within the framework of the Cross-Language Evaluation Forum (CLEF)[1], the Web People Search Evaluation Cam-

[1] http://www.clef-initiative.eu/

paigns (2007-2010)[2] focused on the Web People Search and person name ambiguity on Web pages and aimed at building a system which could estimate the number of referents and cluster Web pages that refer to the same individual into one group. The research was performed on monolingual data.

Another related area is that of detecting the original text over its multilingual versions known as **cross-lingual plagiarism detection** [8]. The goal of plagiarism detection is to find an unauthorized copy of the original textual document in another language. This assumes that there is an initial original text which has been copied. In our research, the goal is to find identical RDF resources referred to the same entity (real-world or figurative). Thus, the original is this entity which is not expressed in some language. The goal of resource interlinking is to find a similarity which will maximize the chance that the two entity representations expressed in different languages refer to the same thing. Thus, there is no "original" in textual sense, all language representations are equal and the entity itself is detached. The "original" is a described entity, and the goal is to find the copies of it in different languages.

In contrast to plagiarism detection, we aim at providing insights into the problem of cross-lingual interlinking given that data are represented in RDF, and we can vary different parameters in order to determine their impact on the interlinking operation. Our goal is to find resources which were created, legitimately and independently, in different languages. It is not important if one data publisher copied or "plagiarized" the resource description from another publisher. Even if it were copied, it would facilitate the process of finding similarity between these resources and, as a consequence, resource interlinking. In plagiarism detection, some modifications made to the original text can also facilitate the detection. However, the difficult part of plagiarism detection is to detect which changes exactly can serve as a proof of plagiarism.

A classification of methods for cross-lingual plagiarism detection is given in [102]. Methods found in the plagiarism detection domain were mostly borrowed from the cross-lingual information retrieval field which is reviewed next.

Another NLP application which deals with information processing in different languages is **Cross-lingual Information Retrieval (CLIR)** [42, 48]. The goal of CLIR is to facilitate information access across languages. This field investigates the ability of retrieval systems to find documents related to a query regardless of the language in which the documents are written. There exist several evaluation tracks. Until 2002, there was a Cross-Language Track at TREC (Text Retrieval Conference). The cross-language retrieval tasks are studied at Cross-Language Evaluation Forum (CLEF)[3] and at NTCIR evalua-

[2] http://nlp.uned.es/weps/weps-3

[3] http://www.clef-campaign.org/

tion workshops[4] (emphasis on Asian languages). Another forum for comparing models and techniques for cross-lingual document retrieval is an Indian Forum for Information Retrieval Evaluation[5][73].

The TIDES (Translingual Information Detection Extraction and Summarization) program promoted the development of language technology which improves translingual information access and correlation. Within this program, evaluation called "TIDES Surprise Language Exercise" has been developed. The participating research groups are presented with a "surprise" language for which cross-lingual technologies should be improved. The challenges to development of translation resources and cross-lingual retrieval for Cebuano and Hindi languages are described in [96]. The main difficulties encountered are the lack of linguistic and textual resources for the Cebuano language and problems with encodings for Hindi [126].

Extended surveys on state-of-the art techniques in the CLIR are provided in [63, 99, 140].

In contrast to our work, the CLIR deals with short queries represented in a natural language, on the one side, and with a big collection of documents on the other side. For RDF resource interlinking purposes, SPARQL queries are used to retrieve the necessary information from a graph. However, such queries are represented in the form of variables and not in the form of textual data which can be compared directly with another text collection. Moreover, the information retrieval system should retrieve related articles per query where *relatedness* can be understood quite loosely. On the opposite, the goal of our research is to identify *identical* resources across data sets and link them using owl:sameAs link the semantics of which is strict, i.e., indicating object equality.

There is a large variety of approaches for tackling multilingualism found across the domains. Figure 3.4 depicts the principal approaches for cross-lingual data processing. The approaches are roughly partitioned into three groups: syntax-based, interlingual, and translation-based. The partitioning has been done according to the nature of the required transformation. The figures below briefly illustrate each of these approaches. Similarity (sim) is computed over (transformed) textual documents.

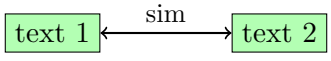


Figure 3.1: Syntax-based methods compare two texts directly.

The next section presents syntax-based approaches which are the simplest

[4] http://research.nii.ac.jp/ntcir/workshop/
[5] http://fire.irsi.res.in/fire

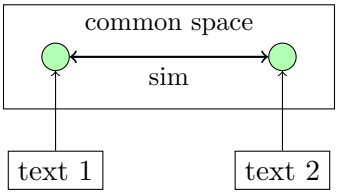


Figure 3.2: Interlingual methods create an intermediate representation for two texts. The comparison is done between the representations.

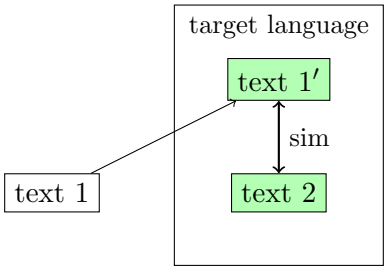


Figure 3.3: Translation methods translate the source language of one text into the target language of the other text.

approaches to deal with two different languages.

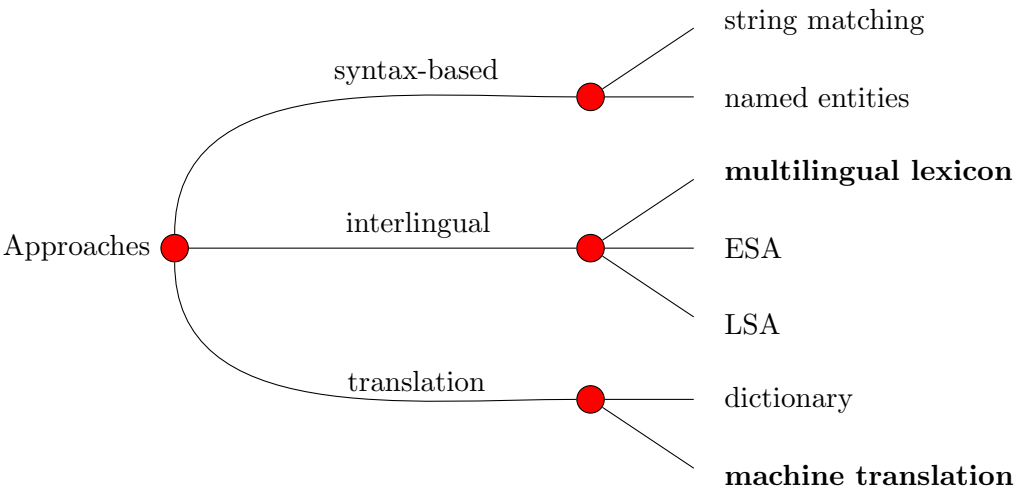


Figure 3.4: General categorization of approaches for processing information in different languages. The methods in bold are experimented with in this thesis.

3.2 Syntax-based Approaches

Syntax based methods rely on syntactic similarities between languages. No additional (external) resources are required for processing texts in related languages.

The methods rely on co-occurrence of common words, n-grams (words or characters), longest common subsequence. So, the more frequent the same elements are in both texts, the more likely the two texts are similar. The major limitation is that the two languages should be alphabetically close. For instance, these methods are not applicable to a language pair such as Arabic - Russian due to different alphabets.

A method for cross-lingual information retrieval using overlapping character n-grams is evaluated in [77]. It is tested across European languages. It is demonstrated that high accuracy can be achieved without applying language-specific resources such as translation. The authors point out that the number of common raw words shared across related languages is less than the number of shared n-grams, so the use of n-grams is more efficient. This method is only applicable to syntactically related languages.

An attempt to improve a bilingual dictionary-based approach by using cognate matching is taken in [74] in the cross-lingual information retrieval. The approach uses cognates, i.e., words which have a common origin, along with approximate string matching techniques in order to improve CLIR. The evaluation is performed on Indian Languages for which queries are in Telugu and documents to be retrieved are in Hindi. The method proceeds as following. First, the query is tokenized into keywords. Then, these query keywords are translated using a bilingual dictionary to obtain the corresponding keywords in a target language. Also, translated query keywords are searched for their corresponding cognates in a target language. Cognate identification relies on the assumption that the likelihood of two words across languages to be cognates is correlated with their orthographic similarity. As cited in [74], the following string similarities are used for cognate identification: the Jaro-Winkler, the Levenstein distance and the longest common subsequence ratio. The query words which have neither bilingual dictionary entries nor cognates are identified and transliterated into the target language. The combined query undergone bilingual dictionary lookup, cognate identification and transliteration is used to retrieve documents in a target language. The approach is evaluated on Hindi news corpora using the 50 Tegulu queries. The authors conclude that the usage of bilingual dictionary together with the cognate identification techniques yield more effective results than using these approaches independently.

A corpus-based translation approach using a Web search engine was adopted in [21]. In particular, online translation service such as Babelfish[6] had been used to translate English queries into Chinese. The untranslated English query words have been used in search engines in order to extract Chinese translations. In experiments with Chinese-Japanese language pair, as Japanese kanji and Chi-

[6]https://www.babelfish.com/

nese traditional characters share the same ideographs, it was possible to use direct mapping through encoding conversion. It was also found that a combination of query translation and cognate matching between Chinese and Japanese performed well.

If two languages cannot be compared syntactically directly, other approaches, more sophisticated, should be considered.

3.3 Interlingual Approaches

Given two different languages, interlingual approaches allow for mapping both languages into a common space independently from each other. This common space is represented by an intermediate layer which, however, contains elements from both compared languages. Interlingual approaches include the use of intermediate representations of both source and target languages. A good example is a multilingual lexicon which contains a fixed set of concepts expressed in different languages. Other methods are based on parallel and comparable corpora and seek to induce patterns in word occurrences. It might turn out that such corpora do not exist for a given domain or the performance of a system trained on a specific corpus will decrease when tested on a more general corpus. Comparable corpora are corpora which contain a pair of monolingual corpora on the same topics described in different languages. However, these corpora are not translations of each other [83]. An example of comparable corpus-based method is the Explicit Semantic Analysis which is discussed below. Approaches which rely on lexical resources often face the problem of scarcity of such resources or the low coverage of terms. However, such resources are indispensable for cross-lingual text analysis and applications.

One type of interlingual mediation is mapping source language terms into a multilingual lexicon. In [125], the EuroWordNet multilingual database is used to find appropriate translations for query terms. For each query in Spanish, the possible EuroWordNet synsets (a set of senses) are identified and then disambiguated using a word-sense disambiguation algorithm. Since each synset has lexicalizations in different languages, the equivalent English lexicalisations of the disambiguated synsets would be used for retrieval against an English document collection. A method which indexes a document collection and a query by EuroWordNet interlingual index is discussed in [43]. This method creates a language-independent representation where each document is represented as a vector of weighted interlingual index records. Document indexing is performed in two steps. First, document terms are mapped to interlingual index records. To do this, part-of-speech tagging is performed and only nouns and verbs are considered. In order to select from multiple synsets, word-sense disambiguation

is applied. Once disambiguated, the synsets are mapped into the interlingual index. The second step includes weighting of this representation, by using standard weighting schemes for example, TF·IDF. Finally, matching between documents shall be performed by computing cosine similarity between document and query representations.

The MLPlag system for plagiarism detection across languages is proposed in [19]. The method analyses word positions and uses EuroWordNet multilingual database for transforming documents into an interlingual representation. EuroWordNet is a multilingual version of WordNet which contains synonym sets (synsets). A unique index or synset identifier is assigned to a synset. The same synset in different languages has the same index which allows for document representation in a language-independent form by substituting a term of a document by an index from EuroWordNet. The evaluation is performed on two datasets in the Czech and English languages. The evaluation on a subset of the JRC-Acquis corpus consisting of European legislative texts [124] achieves the F-measure of 0.72. The authors point out that the insufficient word coverage in a multilingual database of one of the languages can perturb the system performance. The decrease in results can also be due to the topic-specific terms which do not occur in the multilingual database.

Latent Semantic Indexing (LSI) which requires parallel texts in both languages to find co-occurrence patterns is investigated in [72]. LSI assumes the presence of "latent" structure in word usage which is camouflaged by the variability in word usage [11]. The method translates the documents into a language-independent indexing space. The advantage of the method is that it exploits word associations, i.e., contexts in which words appear. Thus, the learned relationships between words can help to retrieve a relevant document even if it does not contain an exact query term. In [22], a parallel aligned corpus is used in 31 languages. The corpus consists of the Bible's translations aligned by verse. The authors found out that much information is contained in inflectional morphemes in morphologically rich languages, thus, text pre-processing might be necessary to improve the retrieval. Overall, a large number of parallel translations in training data improves the precision of CLIR. The problem with using parallel corpora is that it can be quite costly to acquire a large collection of correct translations.

The Explicit Semantic Analysis (ESA) proposed by [38, 39] represents a comparable corpus-based method. This method represents the meaning of a text explicitly via a vector of concepts from Wikipedia. The authors argue that the use of encyclopedic knowledge is the most appropriate medium for programs analyzing natural language texts. The method used an encyclopedia as a set of concepts. Each concept corresponds to the encyclopedic article which contains a

body of text. The method uses knowledge encoded in the text. Given an input text, the method identifies the most relevant concepts by comparing the input text to the text of the articles. The core of the method is that it represents the meaning of an input text using a weighted vector of *all* Wikipedia concepts. A collection of concepts represents an n-dimensional semantic space, and the meaning of each text is a point in this space. The semantic closeness of two texts is determined by their closeness to each other in this space. The document similarity has been computed on a monolingual collection of documents from news. The authors compare the performance of the ESA method to other methods of document representation (WordNet-based, bag-of-words, LSI) and conclude that ESA shows improvements over the current methods described in the literature. Thus, a concept-based text representation is a feasible approach for computing semantic relatedness between documents.

The ESA method can be used for computing similarity across languages. Wikipedia is a multilingual resources as it contains articles on the same topics in different languages. These articles are connected by the interlanguage links. ESA allows for an interlingual document representation where interlingual part is represented by concepts which are described in different languages.

[51] extends ESA by applying it to the problem of cross-lingual semantic relatedness. The ESA concept vector representations are computed on monolingual versions of Wikipedia, however, since concepts are connected via interlanguage links, it is possible to map them and compare. The experiments are performed on cross-lingual *word pairs* in the following data sets: English-Spanish, English-Arabic, English-Romanian, Spanish-Arabic, Spanish-Romanian, Arabic-Romanian. The results on cross-lingual data are lower than on monolingual data. It is also observed that the results improved on languages for which a large collection of Wikipedia articles exists. The authors also compare their extended version of ESA to a machine translation method in which they translate the first word of the input word pair into the language of the second word. Google Translate has been employed to obtain the translations. Once translated, the similarity between words is calculated using the monolingual ESA. Results obtained by the machine translation are slightly lower than the results by the cross-lingual relatedness method. The experiments have been performed on the pairs of words, and the results may be different if longer texts are considered.

The Cross-Language Explicit Semantic Analysis (CL-ESA) has been proposed in [119]. The influence of different parameters of the original ESA mode in the context of cross-lingual information retrieval is studied in [120, 121]. The experiments involved English, French and German Wikipedia articles which exist in all three languages (linked by interlanguage links). The articles have been used in order to construct a common concept space. The experiments have been

performed on two parallel corpora. For evaluation, parallel documents from one language were taken as queries to search parallel documents in another language. Since corpora are parallel, corresponding translations are known. The obtained results using the ESA are compared against the "gold standard" (these known correspondences). The authors conclude that the original ESA settings are plausible though they can be modified so that better results are obtained. However, cosine similarity which defines similarity of query and document vectors remains the best choice.

Compared to the Latent Semantic Indexing (LSI), ESA represents a document in terms of explicit external concepts while LSI computes such concepts from a parallel corpus. In this sense the concepts are "latent", i.e., not explicit but implicit.

Cross-lingual information processing can be necessary for **cross-lingual topic (event) detection** and tracking. The topic "detection and tracking" is concerned with evolution of the event through time [2]. With the growing amount of multilingual information from internet-based sources, internet surveillance systems should be capable to harvest and analyse this information. For instance, data analysts are interested to see how the same news is discussed in different linguistic communities. In cross-lingual settings, a language component which bridges the gap between languages is necessary. The possible solutions include the methods reviewed above. [103] concentrates on automated news analysis and presents a system which tracks news on the same topics in English, German, French, Spanish and Italian. The method, instead of translation, uses several techniques: a) cognates (common strings across languages, including named entities); b) geographical place names mentions; c) mapping document terms to a multilingual thesaurus (thus constructing a vector of identifiers). EuroVoc has been used as a multilingual thesaurus. Each of the thesaurus identifiers has only one translation into several languages, so the document can be represented in a language-independent manner by the identifiers. Text preprocessing (lemmatization, stemming and part-of-speech tagging) is not performed in order to speed up the process. The authors claim that the lack of language normalization does not play a big role, however, if dealing with more inflected languages it might be more useful to perform preprocessing. The authors highlight that the performance is promising but it is worse than the performance on monolingual data. A rich Chinese-English topic corpus which can be used for evaluating cross-lingual topic detection and text analysis methods is introduced in [137]. Once again, topic detection looks for related topics, while data interlinking by owl:sameAs presumes that two objects are identical.

An alternative to an interlingual approach would be a translation-based approach. Such an approach transforms one or both languages directly following

a language model and rules. A translation-based approach can be viewed as a more flexible approach as it intervenes into the language material directly. Depending on the implementation, the approach requires knowledge of language grammar, syntax and usage probabilities.

3.4 Translation-based Approaches

Translation-based approaches are characterized according to the translation means: dictionary or corpus-based approaches, and machine translation (MT). Dictionary-based methods [69] rely on the use of bilingual term lists. Source language terms are substituted by dictionary equivalents of the target language. Thus, the cross-lingual problem is turned into a monolingual one by means of a dictionary. Problems with a dictionary-based method can be the low term coverage, a domain-oriented dictionary which either ignores or consists of specialized terms. Another difficulty lies in the presence of multiple translation variants. In this case, word sense disambiguation is required, otherwise recall might grow at the expense of precision.

Machine translation-based methods rely on a machine translation engine. There are several types of machine translation (rule-based, example-based, hybrid, statistical). With the development of Web, it seems that the statistical machine translation can benefit from it the most. Statistical MT requires training on a large amount of parallel corpora [83]. Given the size of the Web and the large quantity of textual material available online, statistical MT systems can be trained on it. The output translation can be less correct grammatically than that of a ruled-based system, however, it will be compensated by the vocabulary/phrase coverage. The well-known MT engines are Google[7] and Bing[8] translators. Other online translators are BabelFish[9], PROMT[10] and Yandex[11]. The work of [8] extends the classification of [102] by evaluating machine translation method for plagiarism detection purposes.

A method for identifying the same records across databases in different languages is presented in [9]. The proposed method has been evaluated on Japanese image databases where print descriptions are in English and Japanese. However, it can be applicable to other languages as well. The proposed method is based on the comparison of text values of metadata fields, namely, the titles. The prints come from different museum collections (Japanese and Western) and, as a result, the same print can have different titles: a title in Japanese, a title translated

[7] http://translate.google.com
[8] https://www.bing.com/translator/
[9] https://www.babelfish.com/about-us/
[10] http://www.online-translator.com/
[11] https://translate.yandex.com/

into English or latin-transliteration of a Japanese name. To identify the same print across databases in different languages, two representations are used: 1) latin-transliteration of the title and 2) English title. Two types of similarity are calculated:

Similarity based on proper nouns. In English titles, all words which do not appear in a bilingual English-Japanese dictionary are considered as transliterated proper nouns. The degree of similarity grows as the number of matching proper nouns increases. This type of similarity is used to compute similarity between databases where print descriptions are in English or transliterated.

Similarity based on literal translation. In English titles, words which are not proper names are literally translated into Japanese using the bilingual English-Japanese dictionary, and then this translation is transliterated. A degree of similarity is computed between latin-transliteration titles and transliterated versions of the English title's translation.

The precision of proper noun-based similarity (weighting of matching proper nouns along with partial string matching) is 65,4%. The highest precision of 81,4% has been obtained by using literal translation, weighting of matching proper nouns and partial string matching. The authors have not done evaluation between English and Japanese titles expressed in Japanese characters. The proposed approaches have performed well on relatively short pieces of text (record titles).

In the cross-lingual information retrieval, translation can be applied to a query into a target language [41, 75] or to documents into the query language [95]. Dictionary-based approaches [100] employ bilingual machine readable dictionaries in order to replace the source language words by target language translations. The difficulties of such approaches are translation ambiguity, translation of phrases by its compounds, and the low coverage of vocabulary, i.e., unknown modern words might be omitted. In corpus-based approaches, the translation equivalents are obtained directly from parallel or comparable corpora [106].

If translation resources are difficult to obtain for a particular language pair, it is possible to translate both a source and a target language into some third language called a pivot language. This would allow to convert original language representations into a common language representation, so that monolingual similarity methods could be applied.

The present thesis contains experiments involving interlingual and machine translation approaches. As our task is to link RDF resources, we preferred a multilingual lexicon in RDF.

3.5 Multilingual Resources

In order to link RDF resources across languages, interlingual approaches can be applied. As discussed in Section 3.3, multilingual lexicons can be used for creating intermediate representations for entities to be compared. In this section, several multilingual lexical resources are highlighted which are and can be used for bridging the language gap across information in different languages.

Table 3.1 summarizes information about these resources. The resources are organized according to their content. Lexical relations include synonyms, hypernyms, hyponyms, etc.

Table 3.1: Multilingual resources.

resource name	#languages	content	type of relations
JRC-Acquis	21	parallel corpus	paragraph alignment
Europarl	21	parallel corpus	sentence alignment
Wikipedia	291	articles	cross-language links
WordNet	1	synsets	lexical
EuroWordNet	8	synsets	interlingual index
WikiNet	over 100	synsets	interlingual index
BabelNet	271	synsets	interlingual index
Wiktionary	over 170	lexical entries	lexical
DBnary	21	lexical entries	LEMON ontology

The **JRC-Acquis** [124] is a parallel corpus constructed from the European Union legislative documents including obligations, international agreements, etc. The corpus contains texts in 21 languages. Paragraph alignment is available for 190+ language pair combinations in XML format. Moreover, texts are classified according to EuroVoc subject domains. As pointed out in [124], parallel corpora exist for a small number of language combinations, often involving English. Thus, this corpus enriches publicly available lexical resources by providing alignments between rare language combinations such Estonian-Maltese.

Another parallel corpus consisting of extracted proceedings of the European parliament is **Europarl** [64]. The corpus includes utterances of speakers in 21 European languages: Romance (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavic (Bulgarian, Czech, Polish, Slovak, Slovene), Finno-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek[12]. The corpus consists of sentence aligned texts. Each language version is aligned with its English counterpart.

[12]http://www.statmt.org/europarl/

Parallel corpora are important for training statistical machine translation systems as well as for finding co-occurrence patterns across languages.

Wikipedia[13] is a crowd-sourced encyclopedia which contains articles on different topics and named entities. It contains more than 5 millions articles in English. The features which make Wikipedia particularly valuable are its multilinguality and universality (i.e., it is not domain-specific). As per the end of 2015, there are 291 different language editions[14]. Due to the cross-language links, it is possible to have descriptions of the same topic in different languages. Thus, Wikipedia can be exploited as a source of comparable corpora, for instance, for identifying word translations [107].

WordNet [82] is a lexical database of the English language. It is free and publicly available. Words are grouped into unordered synsets (sets of synonymous words) used to express concepts. Synsets are interlinked by means of lexical relations: hyponymy (more specific terms such as "piano" and "saxophone" for a "musical instrument"), hypernymy (more general terms such as "rhododendron" for "azalea"), antonymy (terms opposite in meaning such as polar vs. equatorial), meronymy (part to whole relation such as "eye" for "face" or "toe" for "foot"). A synset can contain a brief definition ("gloss") in the form of a short sentence illustrating the use of the synset elements. WordNet differentiates between types (common nouns) and instances (specific entities, for instance, persons or geographic locations). Thus, a tiger is a type of a cat, John Lennon is an instance of a rock star. Later this initiative has been extended to other languages.

EuroWordNet [132] consists of language-specific wordnets which are linked to the English WordNet. It is a multilingual database which contains separate wordnets with lexicalizations in several European languages (English, Dutch, Spanish, Italian, German, French, Czech and Estonian). The wordnets follow the same structure as WordNet. An interlingual index connects the different wordnets together. The languages are interconnected via this index, so it is possible to find related words in another language. The Global WordNet association[15] promotes the development of wordnets for all languages in the world.

A knowledge-rich lexical resource is proposed in [88, 89]. **WikiNet** is a concept network created automatically by exploiting knowledge from Wikipedia. The nodes of this network are concepts represented by Wikipedia articles and categories. The edges are relations between these concepts which are taken from infoboxes, categories and article texts. WikiNet is a multilingual resource as each concept is lexicalized in different languages. These lexicalizations can be

[13] https://www.wikipedia.org/
[14] https://en.wikipedia.org/wiki/List_of_Wikipedias
[15] http://globalwordnet.org/

accessed through the multilingual concept index. Multilingual lexicalizations are created from the interlanguage links. WikiNet mirrors the structure of WordNet, however it covers named entities better.

The **Universal Networking Language** (UNL) is a formal language for representing and describing the information from natural language texts. It can serve as an interlingua for representing content of a text independently of its original natural language. Information from a natural language text is encoded as a graph in which nodes are concepts linked by labeled edges which stand for relations between the nodes. It can be viewed as a semantic network in which nodes are Universal Words and attributes and edges are UNL relations. The use of UNL for multilingual information processing (retrieval and machine translation) is discussed in [15].

Some lexical semantic resources are also published as linked data. The LIDER project[16] aims at providing interlinked language resources (corpora, dictionaries, etc.) for exploitation in multilingual content analytics across different media resources.

WordNet has been converted in RDF [129]. **BabelNet** [91] is a multilingual semantic network which covers 271 languages in BabelNet 3.0 edition. The nodes of this network are concepts and named entities. The concepts are connected by semantic relations. Each node comprises a set of lexicalizations of the concept in different languages. It integrates several lexical resources such as WordNet, Wikipedia, OmegaWiki, Open Multilingual WordNet, Wiktionary, and Wikidata. It also employed statistical machine translation to get translations for BabelNet concepts which are not covered in resource-poor languages. Thus, BabelNet covers languages (vocabularies) in a balanced manner. BabelNet can be used for interlingual representation of multilingual documents due to language-independent concept identifiers.

DBnary [111, 112] is a multilingual lexicon. It provides multilingual lexical data extracted from Wiktionary[17] in various languages[18]. DBnary contains lexical data such as lexical entries for each word, translations to other languages, word senses, definitions, lexical-semantic relations, and morphological information. The structure of this lexical resource is based on the LEMON model[19]. The extracted data is made available as LLOD (Linguistic Linked Open Data). Linguistic data includes data in Bulgarian, Dutch, English, Finnish, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian, Serbo-Croat, Spanish, Swedish, Turkish, and other languages adding up to 21 languages in

[16] http://www.lider-project.eu/?q=what-is-lider
[17] https://www.wiktionary.org/
[18] http://kaiko.getalp.org/about-dbnary/
[19] http://lemon-model.net/

total.

3.6 Matching in the Semantic Web

In the Knowledge Representation domain, knowledge can be represented at schema and data levels. Schema level refers to the way data are structured and reflects the relations between concepts [10]. Data level contains instances which belong to these concepts. In the description logic, this distinction is also known as Tboxes and Aboxes respectively.

In the Semantic Web, ontologies are used to model a domain knowledge, and data can be described according to an ontology. Ontologies may be heterogeneous because they are modeled independently by different people using different terminologies. Ontologies can be also produced by speakers of different language communities, the multilingual aspect increases the heterogeneity.

Figure 3.5 illustrates two levels at which data interlinking can take place. The upper part of the figure represents two classes from Ontology A and Ontology B. Each of these classes is populated with instances. Ontology classes can also be empty, i.e., do not contain any instances.

The process of finding equivalent classes between two ontologies is called Ontology Matching [32]. A set of correspondences between classes is called an alignment.

The process of finding equivalent instances from two different sources is called data matching [25] or data interlinking [52].

The relation between ontology matching and data linking is discussed in [109]. The authors argue that both domains can collaborate and benefit from each other.

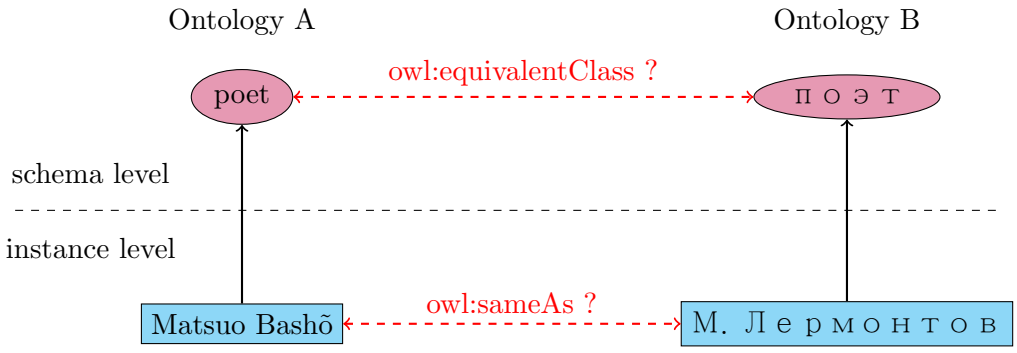


Figure 3.5: Ontology matching at schema level. Data interlinking at instance level. Schema level refers to ontological classes and relations, instance level refers to concrete entities which may belong to some ontology class.

Matching equivalent resources may be done by computer algorithms or humans. The advantage of human effort is that it is possible to find more subtle relations than those of equivalence. The disadvantage is that it can be a slow and labor-intensive process. Due to the growing amount of available data sources which can be quite voluminous, it is more feasible to apply the automatic methods which are capable to provide accurate results.

An overview of methods for ontology matching and data interlinking across languages is presented in the following sections.

3.6.1 Ontology Matching

Ontology Matching (OM) is a widely researched field, and there are many different methods in order to find corresponding classes of properties as discussed in [32]. Many matching methods rely on lexical comparison. This technique is not applicable for matching ontologies expressed in different languages.

In [20], a systematic analysis was done to find the most effective string similarity metric for ontology alignment. This work also explores whether string preprocessing strategies such as tokenization, synonym lookup, translations, normalization, etc. can improve ontology alignment results. The authors mention that preprocessing procedures do not have a strong impact on performance, however they confirm that machine translation improves the results when dealing with different languages. Particular string metrics are suggested depending on the ontologies to be matched. Transliteration is also beneficial if no translation is available.

The Ontology Alignment Evaluation Initiative (OAEI)[20] is a yearly evaluation campaign aimed at comparing the matching techniques and improving the research on ontology matching.

Recent developments have been made in multilingual ontology matching. A MultiFarm benchmark data set for multilingual ontology matching is described in [79]. The creation of such benchmarks is important as it allows for conducting systematic evaluations of approaches. The benchmark consists of seven English ontologies which have been translated into French, Spanish, German, Dutch, Portuguese, Czech, Russian, and Chinese. The ontologies have been aligned manually, thus providing reference alignments. Translation of ontology labels have been performed by humans. The translated version of each ontology is created as a separate ontology expressed in a single language. So, no multilingual labels are present in the same ontology. The preliminary results showed that the best aggregated result of 0.18 F-measure was obtained by the CIDER [45] matching system. CIDER has been executed with default settings. CIDER is a

[20] http://oaei.ontologymatching.org/

schema-based matching system. It also uses a context of input terms, e.g., synonyms, properties, etc. In addition to computing linguistic similarity of terms, it also compares relationships between terms. The systems were not designed for matching ontologies in different languages. The authors also argued that it is important to match structurally different ontologies expressed in different languages. Otherwise, matching systems can use structural information which leads to significant result improvement. Therefore, it becomes more difficult to estimate the influence of multilingual techniques on the system performance. This preliminary evaluation has been extended in [78]. The results obtained by ontology matching systems (which are not designed for dealing with multilingual labels) provide a baseline for the MultiFarm benchmark. Russian and Chinese languages are excluded from the evaluation. This is due to the fact that the evaluated matching systems were not capable to generate alignments between ontologies in Russian and Chinese[21]. The authors conclude that the vocabulary overlap impacts significantly the matching results, however, string comparisons cannot resolve complex correspondences. The best result of 0.31 F-measure has been achieved again by CIDER on the German-English language pair. No particular technique to deal with terms in different languages are used. Overall, the results suggest that the techniques which deal with multilingualism need to be employed.

In 2013, CIDER has evolved into CIDER-CL [44] by including Cross-Language Explicit Semantic Analysis (CL-ESA) (see Section 3.3) into its arsenal. Several languages are supported: English, Spanish, Dutch and German. The results of CIDER-CL on the MultiFarm dataset for OAEI 2013 achieve an average F-measure of 0.17 on matching different ontologies in different languages. The results on the same ontologies were slightly higher with an average F-measure of 0.26; this shows that this schema-matching system is capable to leverage structural information.

A common approach to bridge a natural language barrier consists of transforming a cross-lingual problem into a monolingual one by translating the elements of one ontology into the language of the other ontology [37] using machine translation (see section 3.4). After translation, monolingual matching strategies [32] are applied. In [36, 128, 133], the Google Translate API service has been used. Another way to approach ontology matching is to use external lexical resources. Some of the ontology matching approaches employ Wikipedia's search functionality and interlanguage links for finding mappings [54]. In [71], Wiktionary[22] is used as a lexical background knowledge.

As reported in [29], three systems incorporated machine translation to deal

[21] Cassia Trojahn, personal communication

[22] www.wiktionary.org

with different languages to participate in the MultiFarm track of the OAEI 2014. Table 3.2 shows the best results on different ontologies. AML used Microsoft Bing Translator to translate labels of classes and properties and stored them locally. The translator is queried again if no stored translations are available. AML achieved the highest F-measure of 0.54. XMap++ also used Microsoft Bing Translator. LogMap used Google Translate API.

Table 3.2: Results for MutiFarm track@OAEI 2014. Aggregated F-measure on the task involving different ontologies.

	F-measure
AML	0.54
LogMap	0.40
XMap++	0.35

Table 3.3 shows the highest results for the MultiFarm track of the OAEI 2015 on different ontologies. The given results consider all alignments (including empty and not generated)[23]. AML, XMap, and CLONA employed Microsoft Translator. LogMap used both Google translate and Microsoft translator. LYAM++ takes advantage of the multilingual database BabelNet (see Section 3.5).

Table 3.3: Results for MutiFarm track@OAEI 2015. Aggregated F-measure on the task involving different ontologies.

	F-measure
AML	0.51
LogMap	0.41
CLONA	0.39
XMap	0.24
LYAM++	0.14

One of the reasons why some results are lower than in 2014 is that the test set included more languages such as Arabic, Italian and Russian. As reported by organizers, for some systems it was difficult to deal with all pairs of languages.

Overall, the MultiFarm evaluation shows that, in ontology matching, specific cross-lingual techniques are beneficial. Thus, it is reasonable to test them in the context of cross-lingual instance matching.

Moreover, the MultiFarm evaluation in 2015 shows that systems using Microsoft Bing Translator or Google Translate performed better than systems us-

[23] http://oaei.ontologymatching.org/2015/results/multifarm/index.html

ing BabelNet. However, they have not been tested interchangeably in the same matcher. We address this problem in the experimental part of this thesis.

A method for OM based on linguistic information is proposed in [105]. The method creates "virtual documents" for nodes, thus encoding the meaning of these nodes into a document. It exploits the RDF structure of ontologies in OWL/RDF. Linguistic elements which are in the local description of the node such as the values of rdfs:label, rdfs:comment properties are collected into the virtual document. In addition, it also contains labels from neighboring nodes. To compute similarity between nodes, standard TF·IDF and cosine similarity are applied. As reported in [105], the results on the OAEI 2005 monolingual benchmark tests showed that the virtual document-based method outperforms methods which do not take into account information from neighboring nodes. This especially concerns the test cases when the description of the given node is not sufficient. The virtual document method outperformed string comparison methods as well as a WordNet-based approach (the relatedness of words depends on the distance between them in WordNet). The proposed method of extracting virtual documents uses information from the adjacent nodes of the given node. It does not provide a notion of levels or the depth of graph traversal. Moreover, information from the neighboring nodes comes from triples in which the given node can be a subject or an object.

A machine learning approach for ontology matching in different languages is evaluated in [123]. The machine learning approach using a ranking support vector machines (SVM) is evaluated on financial data in different languages. This SVM ranks good matches higher than the bad ones. 42 features (similarity and structure-based) have been used in the training. All translations have been done using Microsoft Bing Translator. The main conclusion is that the availability of multilingual information (matching across several languages) improves the performance of ontology matching system in both multilingual and cross-lingual scenarios.

Similar findings have been reported in [90] where semantic relatedness between words is computed using a multilingual knowledge-base approach. The method takes two words in different languages and returns a measure of semantic relatedness between them on the basis of information in BabelNet. The authors argue that the joint use of multiple languages improves the performance of the method.

Lexical Hierarchies Apart from ontologies in different languages, there are other hierarchies which can be expressed in different languages and which can be

interlinked. The notion of a knowledge organization system has been developed in library and information sciences. Such a system organizes information by means of controlled vocabularies such as classification schemes, subject heading, taxonomies and thesauri [56]. A vocabulary is a predefined list of terms or short phrases aimed at cataloging information to facilitate its retrieval. Such terms can be used to annotate (tag) digital resources so that they can be retrieved more easily. Thesauri can be used by indexers to apply index terms to text collections. The examples of general-purpose thesauri are Roget's and WordNet which contain sense relations such as synonym and antonym. One of the well-known domain-specific thesauri for describing objects of art and culture is the Art & Architecture Thesaurus (AAT)[24].

Even though RDF entities are often real-world individuals and events, linguistic resources such as thesauri, dictionaries, corpora are also available in RDF. There are many lexical-semantic resources for different languages and domains grouped in the Linguistic Linked Open Data cloud[25] [23] which is a sub-cloud of the Linked Open Data (LOD) cloud[26]. Linking heterogeneous multilingual *linguistic* resources is also an active research area. These linguistic resources should be interlinked to enhance their interoperability and usability [24, 76].

There is quite a number of thesauri published as linked data and thus available in a machine-readable format on the Web. SKOS (Simple Knowledge Organization System)[81] is an ontology widely used for representing conceptual hierarchies on the Web. The Environmental Applications Reference Thesaurus (EARTh)[1] is a SKOS multilingual dataset containing terms related to the environment. Other environmental thesauri available as Linked Data are GEneral Multilingual Environmental Thesaurus (GEMET)[27], EUNIS[28], Geological Survey of Austria (GBA) Thesaurus[29] - a bilingual (German/English) vocabulary for representing geodata. Some of these terminological resources are interlinked, for example, EARTh thesaurus has links to GEMET, AGROVOC as well as DBpedia. AGROVOC[30] covers areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, etc. AGROVOC is available in 23 languages and is aligned with other multilingual vocabularies related to agriculture. The use of English as a common language for labels has been used in order to link AGROVOC to other thesauri

[24] http://www.getty.edu/research/tools/vocabularies/aat/
[25] http://linguistic-lod.org/
[26] http://lod-cloud.net/
[27] http://www.eionet.europa.eu/gemet/en/themes/
[28] http://datahub.io/dataset/eunis
[29] http://datahub.io/dataset/geological-survey-of-austria-thesaurus
[30] http://aims.fao.org/standards/agrovoc/concept-scheme

[86]. EuroVoc[31] is a multilingual, multidisciplinary thesaurus covering the activities of the European Union and is available in 23 EU languages. A multilingual thesaurus for the Social Sciences – TheSoz 0.93 – is presented in [138]. This is a SKOS-based thesaurus containing concepts with labels in English, German and French languages. The HTML representation of the thesaurus is available online[32].

A vocabulary-based approach for matching multilingual hierarchies (ontologies and thesauri) is proposed in [30]. The approach is multilingual in that it exploits all multilingual labels of entities to be matched. The author reports the F-measure of 0.82 for matching AGROVOC–EuroVoc thesauri. This work also confirms that the availability of multiple labels per entity improves the system performance.

With the development of the semantic web, the discovery of information can be largely improved if data publishers provide their data as linked data. However, due to the variety of vocabularies, it become crucial to link one source of data to another. This linking is supported by semantic equivalence statements, e.g., owl:sameAs, skos:exactMatch. Following such links, information about the same entity can be merged from different sources thus favoring the discovery of facts about this entity. The main objective of question answering over linked data [17, 26] is to facilitate, in part, multilingual access to the information originally produced in different culture and language.

3.6.2 Data Interlinking

In the Semantic Web, the task of determining whether two RDF entities from different data sources denote the same entity and can be linked together is known as data linking or instance matching [34]. As there are resources (webpages) and links between them in the Web, so there are resources and typed relationships between them in the Semantic Web. In the Semantic Web, several different URI references can refer to the same entity and the ability to identify equivalent entities is crucial for Linked Data. Interlinking RDF data sets is the process of setting sameAs links between semantically related entities, i.e., entities referring to the same object. The usage of owl:sameAs links has been studied in [27, 28, 58]. To facilitate data integration and knowledge sharing on the Web, interlinking tools capable of handling entities denoted in different natural languages are very important [46].

Nowadays, many data publishers make available their data as linked open data. Apart from DBpedia [6, 14] with its multilingual versions [3] that became

[31] http://eurovoc.europa.eu/
[32] http://lod.gesis.org/pubby/page/thesoz/

a central hub of the Linked Open Data (LOD), the LOD cloud is growing by integrating more and more RDF data. Even though there are many dataset in English in the LOD, datasets in other languages are also published. The work described in [134, 135, 136] shows the initiative of converting Chinese equivalents of Wikipedia (i.e., Hudong Baike[33] and Baidu Baike[34]) into RDF data sets. This effort resulted in a large-scale cross-lingual knowledge base – XLORE[35] [70].

The Quran dataset is presented in [114]. It is a multilingual parallel RDF representation of translations of the Quran in 43 languages including rare languages such as Divehi, Amharic and Amazigh. This dataset is also linked to DBpedia and Wiktionary.

This section focuses on systems which can deal with instance linking cross-lingually.

A well-known evaluation initiative for the evaluation of instance matching techniques and tools is Instance Matching (IM) track at OAEI already mentioned in Section 3.6.1.

IM@OAEI focuses on discovering matching instances in different RDF and OWL datasets. The participants of the track link RDF resources across various datasets. The performance is evaluated by comparing the generated links with the pre-defined reference alignments provided by IM organizers. The generated links (L) are compared against the gold standard, i.e., reference links (R). The metrics and evaluation process described below are widely for evaluating interlinking methods. The performance of each interlinking method is evaluated by means of standard metrics:

Precision measures the correctness of the generated links:

$$Prec(L,R) = \frac{\mid L \cap R \mid}{\mid L \mid};$$

Recall measures the completeness of the generated links:

$$Rec(L,R) = \frac{\mid L \cap R \mid}{\mid R \mid};$$

F-measure is the harmonic mean of precision and recall:

$$F\text{-}measure = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

[33] http://www.baike.com/
[34] http://baike.baidu.com/
[35] http://xlore.org

The evaluation schema is shown in Figure 3.6.

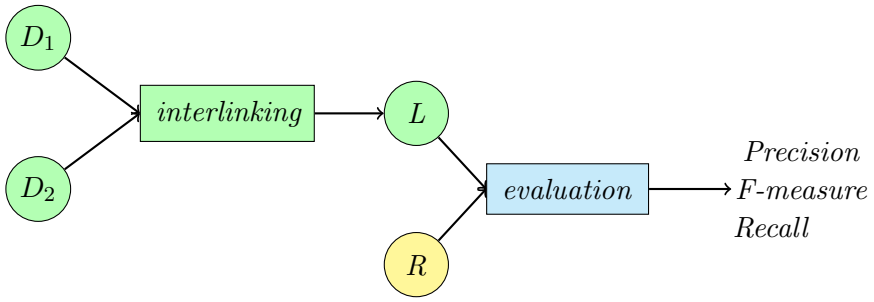


Figure 3.6: Experimental Setup and Evaluation. Two RDF data sets (D_1 and D_2) are input. The interlinking component determines degree of similarity between RDF resources. Obtained links (L) are compared against reference links (R) through precision, recall and F-measure.

The problem of instance interlinking has been studied in many works. Different interlinking approaches have been proposed in the last years. A matching approach which selects RDF predicates using entropy and entity labels is described in [4]. Selection of candidate matches is performed by indexing names of the resources and applying similarities (name and geographic) is used in [94]. In [94], virtual documents were created for resources as the resource descriptions were relatively poor in a source dataset. Both systems have been evaluated in IM@OAEI2011. The evaluation has been performed on monolingual data. To note, candidate selection based on direct string matches between source and target resources is problematic in a cross-lingual context. If, for instance, translation is done before candidate selection, it can be argued that translation on distinct property labels can be not as good as on the same labels taken in context (i.e., assembled into a virtual document).

A time-efficient approach based on the triangular inequality in the metric space for approximating the distance between instances is proposed in [92]. A schema-independent approach is presented in [93]. The approach selects discriminative RDF predicates on the basis of coverage and discriminability. Both systems were evaluated on monolingual data sets. Some approaches use linkkeys i.e., pairs of properties characterizing equivalent resources [5]. A survey on other instance matching systems can be found in [34].

IM@OAEI2014 included two tasks one of which is identity recognition, i.e., the goal is to find instances which refer to the same real-world object. Five systems have participated in the IM track. The datasets contained instances describing books. The test data has been generated by transforming descriptions of the original data. One of the value transformations involved replacing English terms with the corresponding Italian translations [29]. Thus, the proposed task

was cross-lingual instance matching. The best results, an F-measure of 0.56, have been achieved by RiMOM-IM system. RiMOM-IM used Google Translate to translate data into English. Once translated, data preprocessing steps such as stop word removal have been also performed. The method relies on candidate pair generation (a.k.a blocking) which allows to limit the number of candidate instances to be matched in order to avoid comparing all possible instance pairs. Given that each instance is described by RDF triples, this blocking method uses the top n words of the "object" for each predicate as index keys for instances. Thus, an inverted index is generated on the objects. Instances which share common objects are taken as candidates. Different similarity functions are applied to separate predicate values. To compute similarity between instances, similarities over predicates are aggregated into a final matching score. The authors stated that the use of translation helps to improve the results. However, since only strings are used as predicate values for selecting candidate pairs, the relation information between instances is not taken into account [113].

IM@OAEI2015 evaluated six systems participating in the the track. The goal of author disambiguation task was to find instances referring to the same author on the basis of his/her publications. The best results, an F-measure of 0.98, have been achieved by the Lily system[36]. From the reported results, it stays unclear whether any technique had been used or not to deal with different languages.

The absence of standard test cases for the evaluation of cross-lingual instance matching methods represents a significant impediment to the improvement of such methods.

There are existing data interlining toolkits which are capable to deal with multilingualism. RDF-AI, a framework and a tool for RDF data sets interlinking and fusion, is described in [110]. The systems includes several modifiable modules. The preprocessing module incorporates a translation service. The configuration parameters need to be supplied by the user for each step. The system has been evaluated on the works of J.S.Bach from two different datasets. The dc:title property values are translated from German into English using Google Translate API. The authors argue that the highest precision of 95% is achieved due to the translation. As it is reported, the precision drops to 87.3% if no translation is applied.

The Silk link discovery toolkit is proposed in [131]. The tool requires user configuration of the linkage rules in the Silk Link Specification Language. Data access parameters as well as various similarity metrics should be specified. The toolkit includes a $< translateWithDictionary >$ function which allows to translate a string using a provided dictionary file. As it can be expected, substituting

[36] http://islab.di.unimi.it/im_oaei_2015/index.html

strings using a dictionary can lead to a poor quality of such translations and other dictionary-based pitfalls mentioned in the section 3.4.

The LOD datasets as well as the linking tools which facilitate link discovery are mostly concentrated on Western languages. However, there is research supporting data interlinking in Asian languages.

A novel method for matching Chinese, Japanese and Korean LOD resources is discussed in [59]. These three languages share many Chinese ideographs which are collected in the Unihan database. The database contains information about the pronunciation as well as possible variants (number of stokes) for the same ideographs across these languages. The authors propose a new Han Edit Distance which takes into account pronunciation information and the number of different strokes between characters. The evaluation of the method has been performed at character- and word- levels on the word pairs shared by the three languages. The evaluation has been performed against the Levenshtein edit distance which is a widely used string similarity measure. The proposed approach outperformed the Levenshtein distance by 0.25 F-measure for each test case. Though it is not explicit in the paper, the limited use of Levenshtein distance at word-level comparison may be due to the fact that Chinese words are short in length. The proposed method relies on the usage of cognates shared across the languages. The evaluation has been performed at a low level in terms of granularity (words). The method is syntax-based, and the similar attempts for other language pairs have been undertaken as described in the Section 3.2.

Novel methods for computing similarity between Korean words (Phoneme distance) and transliterated Korean words (Transliterated distance) are proposed in [57]. The Phoneme distance relies on the distribution of phonemes across the syllables in order to compute distance between Korean strings. The Transliterated distance takes into account the phonetics of the Korean language.

3.6.3 RDF Resource Representation

As it is shown in Figure 2.5, in an RDF graph, knowledge is partitioned into slots across properties and objects which represent different characteristics of RDF resources. In this way, a detailed description can be created. As mentioned in Section 3.6.1, virtual documents can be used in order to represent resources. However, there are different ways to build those documents via graph traversal. In other words, how to choose a subgraph which represents a particular RDF resource? Several methods for instance extraction, which have been proposed in [49], are reviewed below.

I For a given resource, only *immediate properties* are considered. The disadvantage of such representation is that an important part of related informa-

tion can be missed.

II The type of nodes is taken into account. For a given resource, immediate properties plus the properties of the *blank nodes* connected to this resource are considered. The method is called a Concise Bounded Description. Since the method depends on the presence of blank nodes in an input dataset, its use can be limited.

III The graph is traversed according to a *specified depth* from the given resource. The authors argue that the traversal by two edges forward and one edge backwards from the resource is a good compromise for collecting information. This method is called a Depth Limited Crawling.

Given such resource representations, three methods for finding the distance between pairs of resources are proposed in [49]:

- Feature vector-based measure: the shortest path from the given resource in the RDF graph is mapped to features, and a set of nodes obtainable through this path is mapped to values of each feature. The similarity between two instance is computed on the basis of shared properties.

- Graph-based measure: this similarity relies on the overlap of both nodes and edges between two graphs.

- Ontology-based measure: this measure considers only ontological information attached to the root node.

The creation of virtual documents in the interlinking framework proposed in Chapter 4 is based on the graph traversal according to a specified depth. The proposed cross-lingual string method relies on *literals* in RDF graphs. However, we do not use backward links if it is possible to traverse a graph by moving forward. According to the example in Figure 2.5, the names of the properties in two different datasets can be in different languages in a cross-lingual context. Thus, relying on common paths is not sufficient. In two graphs expressed in different languages, the names of the nodes and the edges are in different scripts, so no overlap is possible. Not all RDF datasets are described with respect to a well-structured ontology. The interlinking framework proposed in Chapter 4 aims at such cases. Moreover, even if two ontologies describe two different datasets, these ontologies should be the same or similar in order to be useful. If ontologies are in different languages as well as data, their use will not facilitate interlinking.

3.6.4 Classifications of Matching Techniques

There are several classifications of matching techniques. Ontology matching techniques are classified in [32]. Instance matching techniques are classified in [34]. These classifications are general and comprehensive. So it is possible to position our research into both of them.

Figure 3.7 shows matching techniques used in our approach according to the ontology matching classification. Initially, depending on the kind of input, the matching is divided into content-based and context-based. Content-based matching uses information which comes directly from the content of datasets to be matched. Context-based matching relies on external sources of information. Our approach manipulates entity descriptions which are found in the graphs themselves. So, it resorts to content-based matching. The content-based matching is further split into terminological, structural, extensional and semantic. Only first two techniques are relevant. We use graph-based techniques in order to navigate RDF graphs. And the matching itself is based on language elements collected from graphs.

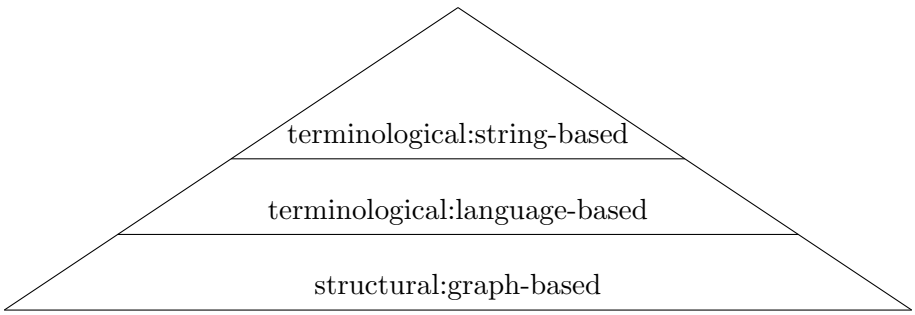


Figure 3.7: Interlinking techniques used in the proposed approach described in Chapter 4. We situate our interlinking approach following the classification based on the origin of information from [32].

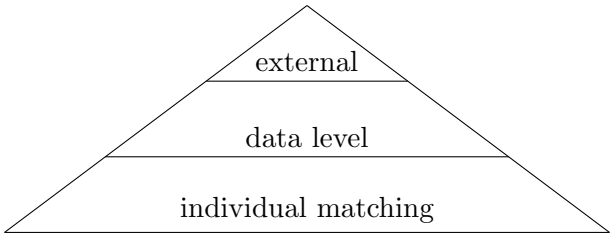


Figure 3.8: Interlinking techniques used in the proposed approach described in Chapter 4. We situate our interlinking approach following the instance matching classification from [34].

Figure 3.8 shows matching techniques used in our approach according to

instance matching classification. With regard to granularity criterion, our approach belongs to individual matching which aims at finding identical entities referring to the same real-world object from different datasets. Data-level methods use information from instance level as in Figure 3.5. Finally, internal techniques make use of information only from datasets to be matched. This notion of internal vs. external techniques corresponds to the notion of content-/context-based matching described above.

Even though the cited classifications can accommodate a vast variety of interlinking methods, they do not necessarily reflect the interdependence of the classified techniques.

3.7 Summary

This chapter reviewed the work on cross-lingual information processing across several research fields: databases, NLP and Semantic Web. Each of these fields faces the problem of object reconciliation, i.e., connecting different representations of the same object together. The evaluations performed in these fields demonstrated the utility of the NLP techniques for detecting identical entities across datasets. There are many methods and techniques which help to overcome the language barrier, however, information access between languages with completely different structures and origins remains a challenging task. There are methods which rely more on orthographic similarity between languages, i.e., matching on character n-grams over languages or using overlap in vocabulary of related languages. These methods are of little use if two languages use different scripts. Overall, little is known of the effectiveness of the linguistic methods in cross-lingual data interlinking. Hence, there is a need for a framework allowing to assess these methods in a controlled environment.

In the next chapter, we propose a general framework that organizes solutions of various nature to deal with cross-lingual data interlinking and allows to compare these solutions.

Chapter 4

General Framework for Cross-lingual RDF Data Interlinking

Abstract. In this chapter, we propose a general framework for cross-lingual data interlinking. It consists of five components including, mainly, RDF resource representation as documents, language normalization, similarity computation and link extraction. The chief component is language normalization as it allows for a homogeneous representation of resources. The framework allows to evaluate cross-lingual techniques in a unified manner.

The previous chapter presented many techniques for transforming two texts written in different languages into some common representation such that similar elements can be detected. In data interlinking, some of these cross-lingual techniques have been tested separately. However, such tests make the comparison of their performance difficult.

This chapter presents a general framework for cross-lingual RDF data interlinking. The framework extends similarity-based data interlinking by accommodating cross-lingual techniques. Therefore, in this thesis, this framework is used to evaluate cross-lingual techniques systematically.

The framework is based on the following underpinning principles:

1. If two URIs denote the same objects, their description should contain common textual elements.
2. If these descriptions are in different natural languages, NLP techniques can be used to bring them in a common space.
3. Once in a common space, some similarities may identify better URIs with common textual elements.

4. The more language data there are, the more accurate this identification will be.

Section 4.1 introduces the overall architecture of the framework. Each component of the framework is presented from Section 4.2 to 4.6. Within this general framework, the cross-lingual string-based approach stresses the importance of textual elements in graphs to be interlinked as well as the availability of language resources. The proposed approach uses declarative knowledge about resources (knowledge asserted in triples). To that extent, we collect all textual information by exploring the neighborhood of an RFD resource within the considered graph. RDF resources are represented as documents consisting from literals harvested from graphs. Once created, these documents go through a language normalization component. Finally, similarity between documents is taken for similarity between resources. Section 4.7 provides an extension to the classification of matching techniques described in the previous chapter. It suggests that the requirement for external resources grows as the languages to be analyzed differ from each other.

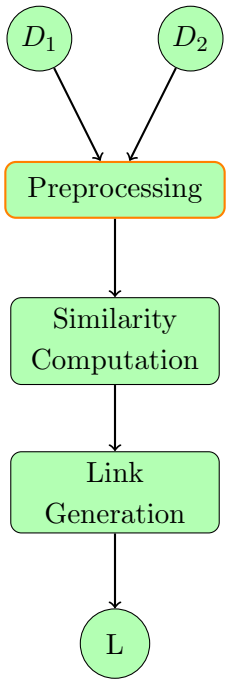


Figure 4.1: The general scheme of similarity approaches to data interlinking.

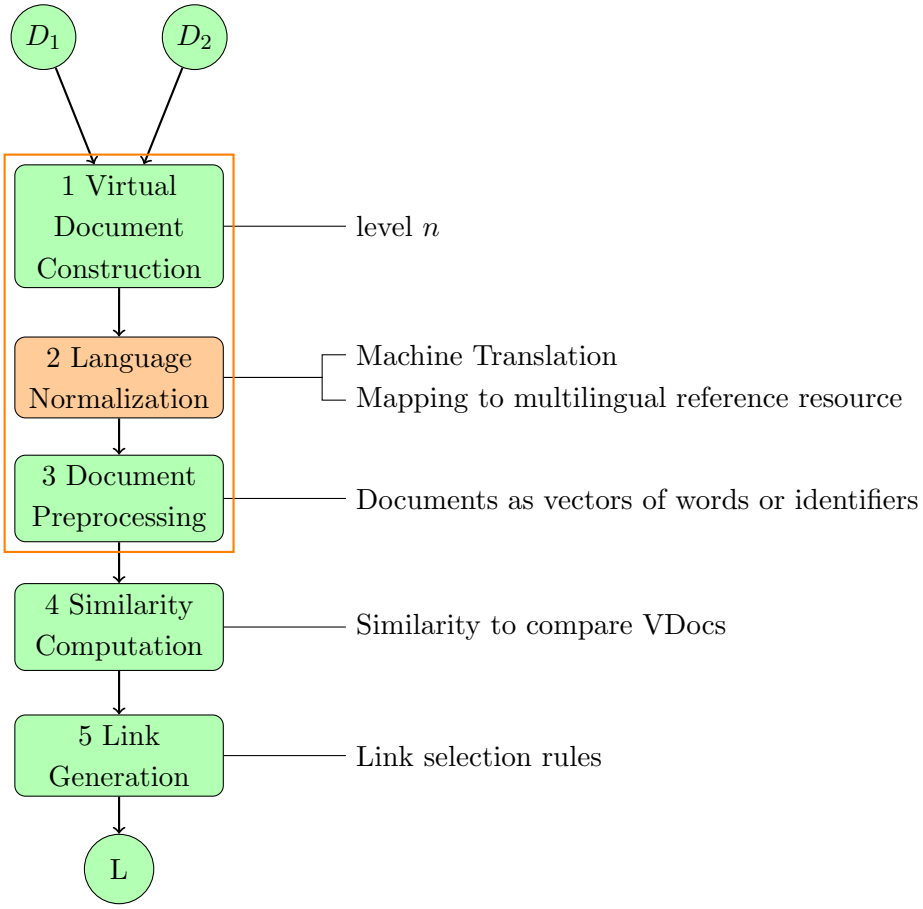


Figure 4.2: Framework for cross-lingual RDF interlinking.

4.1 Overall Architecture

The proposed approach belongs to the family of similarity approaches. Similarity approaches to interlinking consist of several main components which are depicted in Figure 4.1. Figure 4.1 displays the general framework of data linking by similarity. RDF data are an input. The data are normalized using preprocessing techniques, and similarity is computed (Similarity Computation) between normalized RDF data. The links (Link Generation) are extracted on the basis of similarity.

This general scheme is extended in order to accommodate cross-lingual RDF interlinking. The framework for cross-lingual RDF data interlinking is presented in Figure 4.2.

The proposed framework includes five steps:

1. Construct virtual documents: given two data sets with a resource representation in different natural languages, extract language data for each

RDF resource. Thus, a "virtual" document is created for each resource. The idea of creating a "virtual" document has been employed in ontology matching introduced in the section 3.6.1.

2. Normalize languages: If resources are described in different natural languages, it is necessary to find ways to access to their meaning despite their differences in forms: language unification methods are necessary in order to make these languages comparable computationally.

   We achieve language unification by projecting the vocabularies (texts from virtual documents) into the same space. This is the space in which language elements can be comparable. We explore two such spaces:

   - a space of words (strings) created by applying MT on the source documents;
   - a space of identifiers created by mapping texts from virtual documents to a multilingual lexicon.

3. Preprocess documents: use standard document cleaning techniques in order to prepare documents for similarity computation.

4. Compute similarity: compare virtual documents in pairs from both sets and find the similarity between two representations of the resource.

5. Extract matches: set an owl:sameAs link between the two most similar representations.

The framework allows to alter different parameters. At the **data level**, RDF data can vary in the following aspects:

- Distinct language pairs;
- Exploring nature of instances: instances can be homogeneous (belong to one ontological class (if any)) or be heterogeneous (mixed). Instances can also identify named entities (e.g., musicians) or generic nouns (thesauri concepts).

At the level of **techniques**, the general framework allows for the following parameters:

- Different similarity metrics (weighting schemes such as TF, TF·IDF, term occurrence; cosine, Jaccard);
- Different extraction algorithms (greedy, Hungarian);
- Different machine translation tools;
- Strategies that do not depend on translation technologies (e.g., mapping to BabelNet).

Each step of the framework is detailed below.

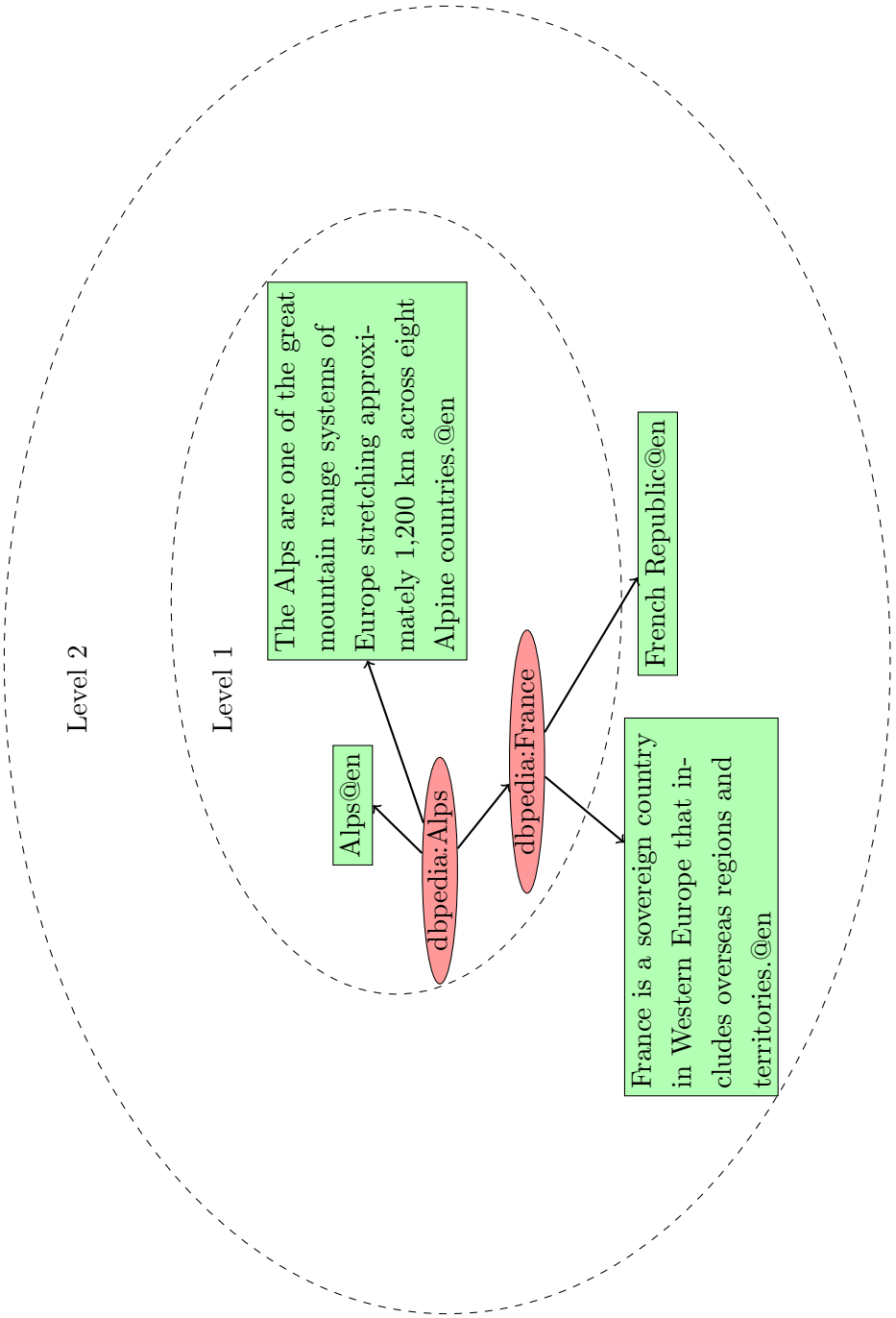


Figure 4.3: Creation of Virtual Documents by Levels. The resource (the Alps) is described in English.

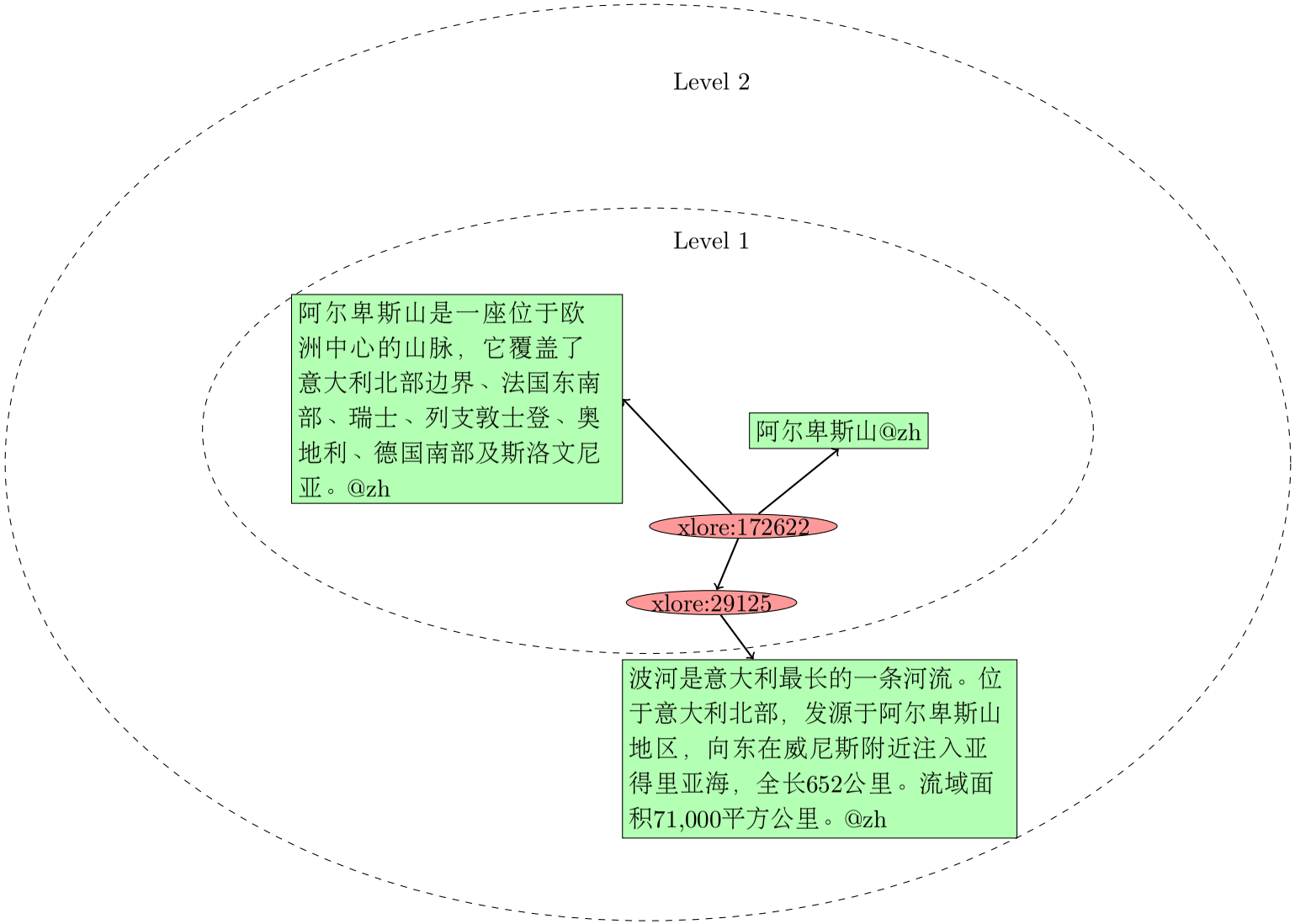


Figure 4.4: Creation of Virtual Documents by Levels. The resource (xlore:172622) is described in Chinese.

4.2 Virtual Document Construction

A British linguist J.R.Firth [35, p.11] wrote that "You shall know a word by the company it keeps" pointing out the important role of a lexical context while analyzing a meaning of a word. We can rephrase this expression into "You shall know an RDF resource by the company it keeps". The company of an RDF resource are all labels which appear in its neighborhood. These labels constitute the context produced by data publishers.

The resources are represented as Virtual Documents in different natural languages. The intuition of converting a graph into a document representation is that even though the taxonomy (structure) of graphs can be similar, the possibility to distinguish between two different things and identify the identical ones relies on their label comparison. Thus, it is important to take into account lexical elements in a graph.

The triples of an RDF graph can have simple strings (literals) as an object which serve as a descriptor for a subject. If the object is a literal, it is stored into a virtual document. If not, the algorithm proceeds to the next URI until it collects all lexical content within a given distance. The lexical content is retrieved when the given resource is in the subject position. We accumulate all the language information for each resource. The purpose of this extraction is to form a virtual document which contains up to n levels of language information depending on the specified distance of graph traversal, see Figure 4.3. The language elements attached to a particular type of relationships are taken into account. The property names are not considered. Resources from the same dataset are described using the same set of properties. Thus, if the same property name appears in many resources, it will not be discriminating. The lexical elements are collected for each level separately and after the Language Normalization stage are concatenated. This is done in order to avoid translating the same information twice at stage 2, as the levels are nested. The performance of the method may depend on the amount of text and discriminative power of labels. Both datatype and object type properties are followed in order to traverse a graph. However, the created virtual documents contain only datatype property values as defined in Section 2.1.

For instance, such properties as "rdfs:label" and "rdfs:comment" usually contain textual data. As an illustrative example, consider Figure 4.3 and Figure 4.4. Figure 4.3 shows a resource description in the English language: literals in a graph are annotated with an English language tag "@en". In this graph, the subject is "dbpedia:Alps" which has for name "Alps" and a comment in English language. Figure 4.4 shows a resource description in the Chinese language. The examples of virtual documents created for each level are presented below.

An example of Virtual document in English for level 1

```
Alps
The Alps are one of the great mountain range systems of Europe
stretching approximately 1,200 km across eight Alpine countries.
```

An example of Virtual document in English for level 2

```
Alps
The Alps are one of the great mountain range systems of Europe
stretching approximately 1,200 km across eight Alpine countries.
French Republic
France is a sovereign country in Western Europe that includes
overseas regions and territories.
```

An example of Virtual document in Chinese for level 1

```
阿尔卑斯山
阿尔卑斯山是一座位于欧洲中心的山脉，它覆盖了意大利北部边界、
法国东南部、瑞士、列支敦士登、奥地利、德国南部及斯洛文尼亚。
```

An example of Virtual document in Chinese for level 2

```
阿尔卑斯山
阿尔卑斯山是一座位于欧洲中心的山脉，它覆盖了意大利北部边界、
法国东南部、瑞士、列支敦士登、奥地利、德国南部及斯洛文尼亚。
波河是意大利最长的一条河流。位于意大利北部，发源于阿尔卑斯山地区，
向东在威尼斯附近注入亚得里亚海，全长652公里。流域面积71,000平方公里。
```

Given a resource, it is possible to collect textual descriptions for this resource following two scenarios:

- Textual description is extracted from one graph;
- Extra textual data can be extracted using federated queries to other datasets (i.e., by looking up URIs from a given resource).

In this thesis, resource descriptions always come from one dataset. No federated queries are used.

4.3 Language Normalization: Machine Translation or Mapping to Multilingual Reference Resource

Once the virtual documents are created, it is necessary to make them comparable, i.e., to project them into the same space in which there exist a similarity. This thesis explores two strategies in particular:

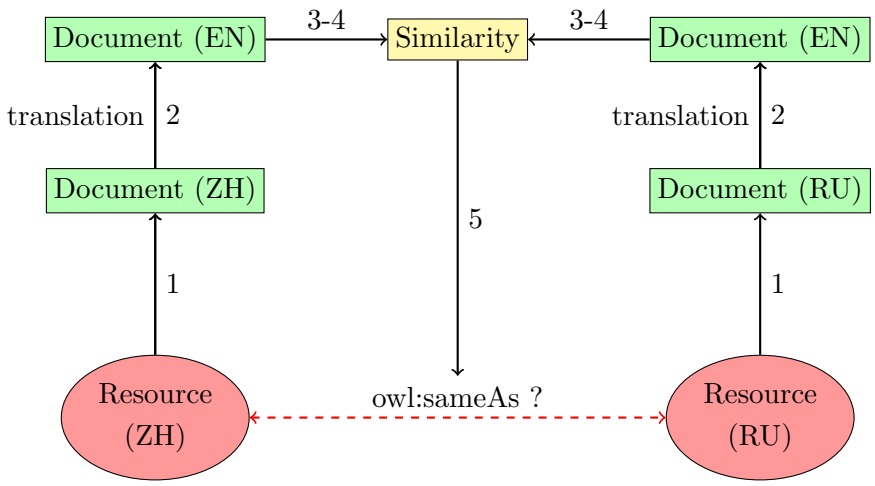


Figure 4.5: Linking Process. Resources are described in Chinese and Russian languages and then translated into English.

Applying machine translation Virtual documents in two different languages are translated using machine translation in order to transform documents into the same language. At this step, virtual documents in one language can be translated into the other language and vice versa or both languages can be translated into some third language. There are several machine translation systems available, see Section 3.4. The choice of translation techniques can also depend on the language combinations, for example, for rare languages, for which there does not exist enough parallel corpora, dictionary-based approaches might help.

Machine translation is used as a black box, only a source and target languages are specified. MT produces one translation which is used.

Figure 4.5 shows an interlinking process where original documents in two languages are translated into a pivot language (English). The numbers on the arrows correspond to the framework's stages described in Figure 4.2.

Mapping to multilingual reference resource An alternative approach is to use Multilingual Resource Mapping instead of translation. For instance, a multilingual lexicon serves as a basis for resource comparison. Document terms are replaced by identifiers from a multilingual lexicon in order to project the words of each language onto the same semantic space. At this step, we represent original documents as vectors of identifiers (IDs). A corresponding identifier (ID) is retrieved for a term. An identifier stands for a sense of a term and very often there are many senses (IDs) per term. If more that one sense exists, word sense disambiguation techniques shall be applied in order to select the best sense. The terms which cannot be mapped in the multilingual resource are discarded and

we do not work with them in our experiments. Mapping to multilingual lexicon can improve recall in cases where the same concept is lexicalized differently: a synonym of this word is used in the other language. To illustrate, suppose there are two virtual documents. The English virtual document contains a word "cat" (domestic animal), while the Russian virtual document, instead of a normal form "к о т" or "к о ш к а", contains a word "к о ш е ч к а" (a diminutive of a "cat"). So, both lexicalizations would be resolved to the same identifier, thus the same idea will be preserved even though it is expressed differently on the surface. To compute semantic relatedness, multilingual reference resources can be used, e.g., BabelNet or DBnary (see Section 3.5).

4.4 Document Preprocessing

Once the terms are translated or replaced by the identifiers, the documents undergo data preprocessing. Preprocessing refers to the processing before computing similarity. Comparable virtual documents are treated as "bag-of-words" following the Information Retrieval paradigm. Different standard NLP preprocessing techniques (transform cases into lower case, tokenization, stop word removal, etc.) are performed at this stage. If documents contain identifiers, these techniques are omitted. For instance, stemming can be useful because it helps to map different surface forms into one, e.g., link, linkage and linking would be reduced to link. Thus, the same essential content is expressed only with one surface form. A well-known stemmer for the English language is a Porter's stemmer [101]. Stop-words, i.e., functions words such as "and", "the" , "of" are not significant and can be removed without harming the entity representation. Once the documents are preprocessed, a vector space model [108] is used to represent terms in a "virtual" document as vectors of features. Virtual documents are represented as vectors of words or identifiers weighted using various weighting schemes for selecting the discriminant words, for instance, Term Frequency (TF) and Term Frequency·Inverse Document Frequency (TF·IDF). Term weight can be assigned by computing $termfrequency$ in a document or distribution of terms across a collection of documents known as $inverse\ documentfrequency$ (IDF). Terms that appear in few documents can be discriminative with regard to the rest of the documents. TF·IDF is widely used in vector space models. The translation process often changes the word order of the original sentences. Being a set metric, TF·IDF is an appropriate weight for such cases.

4.5 Similarity Computation

At the Similarity Computation stage, after transformation of "virtual" documents into vectors, a similarity method should be applied. Similarity between documents can be taken for similarity between resources. The output of this stage is a set of similarity values between pairs of virtual documents. These similarity values are an input for the Link Generation stage. There are many techniques to compute vector similarity. A broad overview of them is given in [32]. Two similarity measures are used for comparing two vectors. Cosine measures the angle of two numerical vectors and is maximal (=1) if two vectors are identical. The Jaccard similarity measures term overlap. The general rule is that the higher *sim (x,y)*, the more likely that x and y denote the same RDF resource.

4.6 Link Generation

At the Link Generation stage, an algorithm extracts links on the basis of the similarity between documents. There are different methods to extract alignments. A broad overview is given in [32]. We use Hungarian and greedy algorithms to extract links. These two methods aim at extracting one-to-one matches. The Hungarian algorithm [87] computes the maximal weight one-to-one matching, while the greedy algorithm computes only a stable local optimum. These are classical methods for link extraction.

4.7 Extension to Classification of Matching Techniques

This section presents an extension to the classification of matching techniques described in Section 3.6.4. This extension concerns the language normalization component of the proposed framework. The instance matching techniques include internal and external techniques which use either information from datasets to be matched only or additionally employ external sources. We pointed out that the classification does not necessarily reflect the interdependence of these techniques. However, when dealing with interlinking data in different languages, the choice of techniques can depend upon the languages to be processed.

Figure 4.6 illustrates an interlinking process. D_1 and D_2 represent RDF datasets to be interlinked. L_1 is the natural language used for describing RDF resources in the D_1. L_2 is the natural language used for describing RDF resources in the D_2. $L_2 - L_1$ denotes the transformation of one language into the other. M stands for matching component and L refers to the resulted links. The external component – here machine translation (or another procedure) – comes before

matching. The external component is essential in particular for languages which use different scripts, e.g., Russian and Hindi. For the related languages, this external component can be omitted, and the matching stage can be performed directly after label preprocessing. In case of related languages, techniques such as vocabulary overlapping can be successfully applied as discussed in Section 3.2.

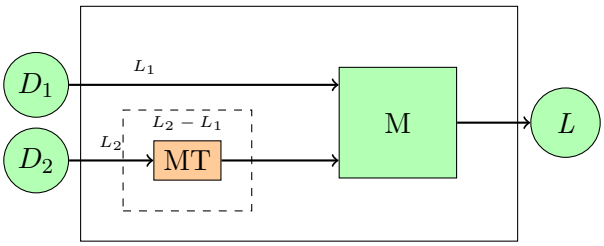


Figure 4.6: Cross-lingual data interlinking process using external resources. In this example, the resource is machine translation (MT).

Figure 4.7 adds to the cited classification w.r.t. cross-lingual instance matching. It suggests that the requirement for external resources grows as the languages to be analyzed differ from each other.

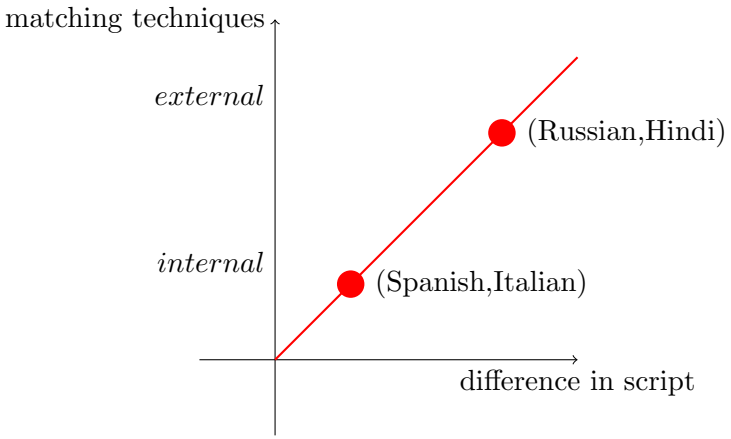


Figure 4.7: The requirement for external resources grows with the dissimilarity of languages to be matched.

There are many techniques and approaches for data interlinking. Linking data across languages constitutes a part of data interlinking task which requires special attention to the language aspect. Even though the current approaches incorporate mechanisms to overcome language differences, there is still a lack of evidence how a particular technique can be beneficial or limited, and which one performs better.

4.8 Summary

Chapter 2 illustrated that two RDF graphs can contain the same knowledge expressed in different languages which are dissimilar orthographically and structurally. Linking these two graphs together requires application of language-specific approaches. Our research goal is to provide and evaluate such approaches.

This chapter introduced a general framework for cross-lingual data interlinking. The major components of this framework include language-specific tools such as machine translation and multilingual reference resources. The important function of the framework is to identify various replaceable components that can be parameterized. Hence, this helps to evaluate cross-lingual linking approaches systematically.

Experiments evaluating the benefits of different parameters are presented in the remainder of this thesis. The following chapters focus on different aspects and describe experiments designed to evaluate the cross-lingual linking techniques.

In the next chapter, we will use this framework to evaluate a machine translation approach. In Chapter 6, the focus shifts to an interlingual method based on a multilingual lexicon. In both chapters, RDF resources represent named entities. Hence, instead of named entities, Chapter 7 considers application of machine translation to resources representing thesauri concepts.

Chapter 5

Linking Named Entities Using Machine Translation

Abstract. In this chapter, we evaluate the suitability of a machine translation approach for interlinking RDF resources. The resources represent named entities and are expressed in English and Chinese. The best F-measure over 0.95 can be achieved by collecting literals from the closest neighbors with minimal preprocessing. The results demonstrate that translating labels is beneficial for resource interlinking, however, the results can vary due to other parameters.

The previous chapter introduced the framework which encompasses language normalization and other parameters for cross-lingual data interlinking. Nowadays, due to availability and advancement of machine translation systems, machine translation became a straightforward approach to deal with information written in different languages. This chapter evaluates the efficiency of machine translation on linking RDF resources. Machine translation instantiates the language normalization component of the framework, it is also the main component as other parameters are applied on its output.

This chapter presents four experiments. Each experiment builds on the previous one by modifying some parameter (RDF data, term weights, link extraction methods). The main experiment is presented in Section 5.1. It introduces a translation-based method which is applied to resources labeled in English and Chinese. According to the general framework, resources are represented as text documents, and similarity between documents is taken for similarity between resources. Documents are represented as vectors using two weighting schemes, then cosine similarity is computed. The results demonstrate that machine translation and the classical Information Retrieval (IR) vector-space model are suitable for interlinking RDF data. The remaining three experiments show how the quality

of generated owl:sameAs links can be impacted by modifying parameters. In Section 5.2 the setting is complexified by adding non-matching entities into an RDF dataset. Sections 5.3 and 5.4 describe an attempt to further improve the results on the most difficult setting by modifying a term weight or using n-grams.

5.1 Experiment I: Original Method

5.1.1 Translation-based Interlinking

The entire data flow with modifiable parameters is illustrated in Figure 5.1.

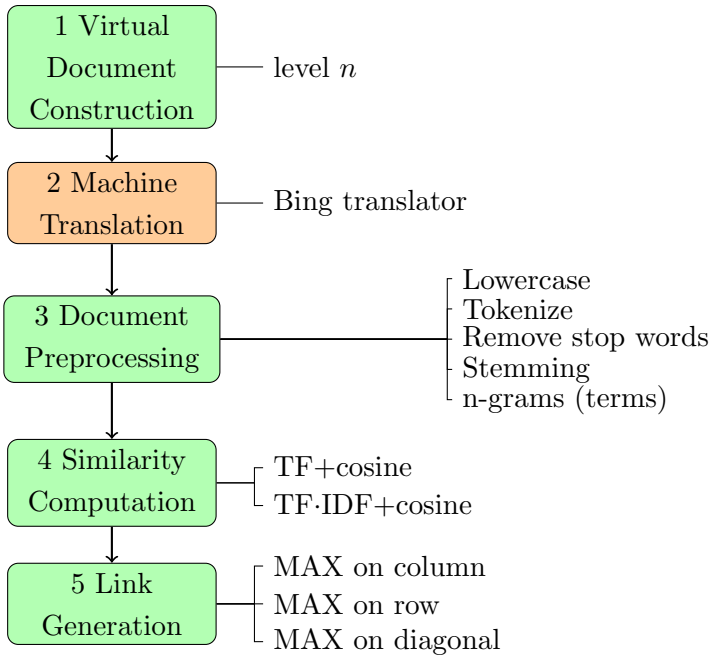


Figure 5.1: Data Flow for Resource Interlinking

Given two RDF data sets, we proceeded as follows.

First, the resources are represented as **Virtual Documents** in different natural languages. To obtain these virtual documents per resource, we collect literals according to the specified graph traversal distance, as described in section 4.2 of Chapter 4.

Next, to make these documents comparable, we use **Machine Translation**.

Once translated, the documents undergo **Data preprocessing**. We constructed four pipelines so that the number of processing steps is growing with each pipeline.

1. Pipeline 1 = Transform Cases into lower case + Tokenize;
2. Pipeline 2 = Pipeline 1 + Filter stop words;

3. Pipeline 3 = Pipeline 2 + Stem (Porter);
4. Pipeline 4 = Pipeline 3 + Generate n-grams (terms, max length = 2).

In order to compute similarity between the resources, we need to compute similarity between the documents that represent these resources. At the **Similarity Computation** stage, we use two weighting schemes: Term Frequency (TF) and Term Frequency·Inverse Document Frequency (TF·IDF) and applied the cosine similarity. The output of this stage is a similarity matrix. The matrix is such that the virtual documents in the original language are on the vertical axis and the translated documents are on the horizontal axis.

At the **Link Generation** stage, the algorithm extracts links from the similarity matrix.

We study three ways of extracting links:

1. We select the best original resource for a translation (selecting the maximum value in a column only);
2. We select the best translation for an original resource (selecting the maximum value in a row only);
3. We select such a translation for which the best original document has this translation as best translation (selecting the maximum value in a column and a row).

5.1.2 Experimental Setup

Our goal is to evaluate how the method described above works and which parameters are important. We also evaluate the suitability of Machine Translation for identifying identical resources.

We would like to observe the effect of the size of virtual documents, preprocessing steps and weighting schemes (TF and TF·IDF) on the results. Basically, we seek an answer to the question: what is the combination of parameters that produces the highest results and can assure the correct match in the interlinking process?

Original RDF Data Sets

The experiment has been conducted on two separate RDF data sets with resources represented in English and Chinese natural languages respectively. Thus, the data consist of the English and Chinese part.

To fulfill the English part, we downloaded the following datasets from DBpedia 3.9[1]: Categories (Labels), Titles, Mapping-based Types, Mapping-based Properties, Short Abstracts, Extended Abstracts. For the Chinese part, we used

[1] http://wiki.dbpedia.org/Downloads39

Table 5.1: Statistics about the RDF Datasets

	# of classes	# of instances	# of properties	# of triples in total
DBpedia	435	3,220,000	1377	72,952,881
XLore	N/D	262,311	6280	7,063,975

a part of XLore[2]: Abstracts, Reference Links to DBpedia, Inner Links, External Links, Infobox Property, Related Items, Synonyms. XLore is the Chinese knowledge-base described in the section 3.6.2.

All the data files have been accessed via a Jena Fuseki server and its built-in TDB store[3]. Statistics of data loaded into triple stores is presented in Table 5.1. Information about XLore classes was not available.

Test RDF subset

We restricted our experiment to five entity types: Actors, Presidents, US Presidents, Sportsmen, and Geographical places. We deliberately included unrelated types in order to observe the difference in similarity within and across types. All entities represented named entities (proper nouns).

The Chinese data has already been linked to the English version of DBpedia and we used a list of owl:sameAs links as our reference link set at the evaluation step. Out of the reference link set provided by XLore, we randomly selected 20 instances per category (Actors, Sportsmen, etc.) for which the two linked resources had text in their properties (more than just rdfs:label). In the US Presidents category, there were only 16 linked instances with text, this was compensated by adding four extra presidents into the category of Presidents. We selected entities that appeared in a reference link set and contained textual information at both levels and in both languages. The result of this selection is a relatively clean corpus which contains textual description of resources at both levels. This allowed us to test the level at which the performance is better.

This provided 100 pairs of entities potentially generating 10,000 links. This RDF data set of 100 resources in one-to-one correspondence is referred to as **Original set** in other experiments described in this chapter.

Protocol

The evaluation was carried out according to the following protocol:

- Build the two sets of resources;

[2]http://xlore.org/index.action

[3]http://jena.apache.org/documentation/serving_data/

Table 5.2: Experimental parameters

VDocs 2	Pipelines 4	Translation 1	Weight 2	Similarity 1	Link Extraction 3
Level 1 Level 2	Pipeline 1 Pipeline 2 Pipeline 3 Pipeline 4	Bing: ZH→EN	TF TF·IDF	cosine	MAX on column MAX on row MAX on column and row

- Run a method configuration and collect the links;
- Evaluate links against the reference links through precision and recall.

5.1.3 Evaluated Configuration

The parameters evaluated are presented in Table 5.2. Thus, 48 settings have been explored in total.

Translate ZH into EN

Once we collected a fixed number of entity pairs for each category in the English and Chinese data sets, we needed to make these entities comparable. For this experiment, we used the statistical translation engine Bing Translator API[4] to translate Chinese virtual documents from the Chinese Simplified into the English language. Sometimes the large documents could not be translated in their entirety, in this case we left everything as is, taking only the part of text that has been translated. It would be interesting to translate documents from English into Chinese as well but RapidMiner does not support Asian languages, so at this point we were dealing only with translations from Chinese into English.

Data Preprocessing and Similarity Computation

The pipelines were designed using the RapidMiner[5] toolkit. We were using RapidMiner 5.3.013 with the text processing extension.

Each data preprocessing step corresponds to a particular operator in RapidMiner. For some operators we can specify parameters. Below are the parameters used:

- Tokenize: mode: non-letters (i.e., non-letters serve as separators between tokens. Because of this, all dates are not preserved in documents);
- Filter Stopwords (English): built-in stopword list;
- The type of weighting scheme (TF or TF·IDF) was set for each pipeline;
- For computing similarity, we were using Data to Similarity Data operator with cosine similarity.

Link Generation

[4] http://datamarket.azure.com/dataset/bing/microsofttranslator
[5] http://rapidminer.com/products/rapidminer-studio/

The output of the similarity computation is a matrix of compared pairs with a value. The 10,000 (100×100) comparisons were tabled as a similarity matrix for evaluation for each tested method. The matrix is such that the vertical axis represents the English DBpedia entities while the horizontal axis represents entities from the Chinese XLore base.

5.1.4 Results

The obtained results are displayed in Figures 5.2 and 5.3. They show that with TF·IDF/ Level 1 we are able to identify more than 97% of the identical entities. The comparison of virtual documents was done at two levels. The results across and within categories using TF·IDF show the same pattern: the best accuracy is achieved at Level 1 and the results get worse at Level 2. The results for TF were systematically lower than those of TF·IDF so we do not report them here.

The similarity of resources within categories is presented in Figure 5.4. Black squares are 5 categories. The similarities are highlighted according to their value, and the color intensifies as the value grows:

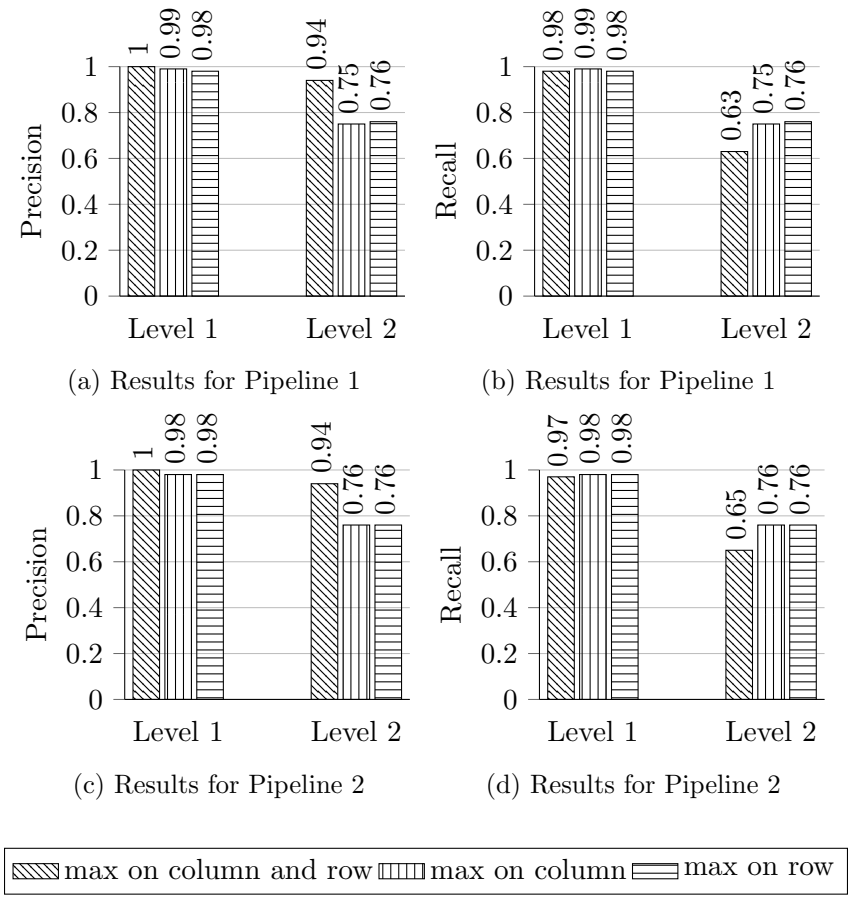
- Values between 0.00 and 0.11 - are suppressed and seen as a white space;
- Values between 0.11 and 0.15 are in light yellow;
- Values between 0.15 and 0.25 are in dark yellow;
- Values between 0.25 and 0.35 are in orange;
- Values between 0.35 and 0.45 are in light red;
- Values between 0.45 and 1 are in dark red.

The correct match is always on the diagonal and the possible confusions are more likely within a category (see the last square which is a "US Presidents" category). This is expected since entities of the same type will have much information in common.

Discussion

The main lessons of this experiment are:

- Our results show the suitability of Machine Translation for interlinking multilingual resources;
- TF·IDF outperforms TF;
- The addition of preprocessing steps seem not to influence the results significantly. The maximum standard deviation is less than 2 points for both precision and recall;
- The quantity of information at Level 1 is usually enough to find a correct match;



(a) Results for Pipeline 1

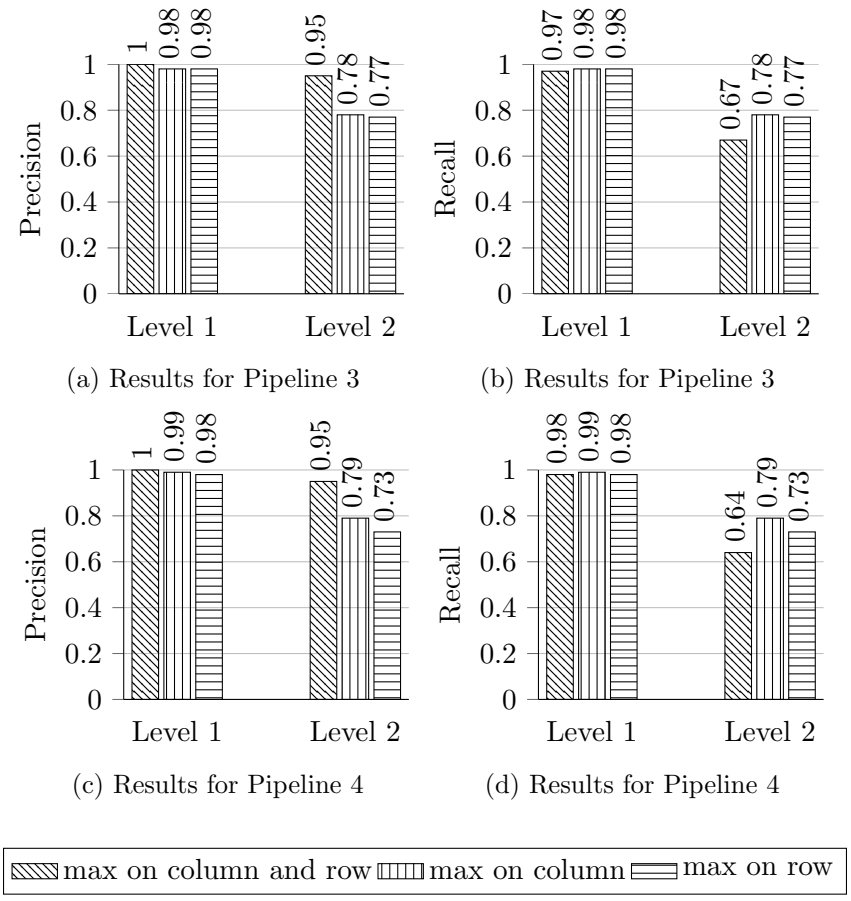
(b) Results for Pipeline 1

(c) Results for Pipeline 2

(d) Results for Pipeline 2

Figure 5.2: Results for Level 1 and Level 2 using TF·IDF

- In general, the results at Level 2 had lower F-measure. This may be explained by supposing that the further we go from the node, the more general becomes the information. If there are many shared properties, then at some point many resources will have the same information (this can be due to the structure of the RDF data set). The discriminant information is thus "diluted" and it becomes harder to detect correct correspondences;
- If there is not enough data at Level 1 then by collecting information from Level 2 it is possible to improve the results. This gives us an intuition that the necessity of proceeding to the next level from Level 1 depends on the amount of data at Level 1. We saw this with one of the error cases when comparing across categories.



(a) Results for Pipeline 3

(b) Results for Pipeline 3

(c) Results for Pipeline 4

(d) Results for Pipeline 4

Figure 5.3: Results for Level 1 and Level 2 using TF·IDF

5.2 Experiment II: link extraction by Hungarian and Greedy methods

In this section, the results are obtained by means of other link extraction methods: the Hungarian and Greedy. We describe results on the Original set mentioned in the previous section. The experimental parameters remained the same as described in the Experiment 5.1, though only TF·IDF and cosine are applied since TF showed lower F-measure previously. Moreover, an extra dataset is added and referred to as Original set + noise. **Original set + noise** contains 10 additional entities in each language side which do not have a match in the other language. This has been done in order to observe how similarity works when entities do not have matches. The results on the Original set are presented in Table 5.3. The results on the Original set + noise are presented in Table 5.4.

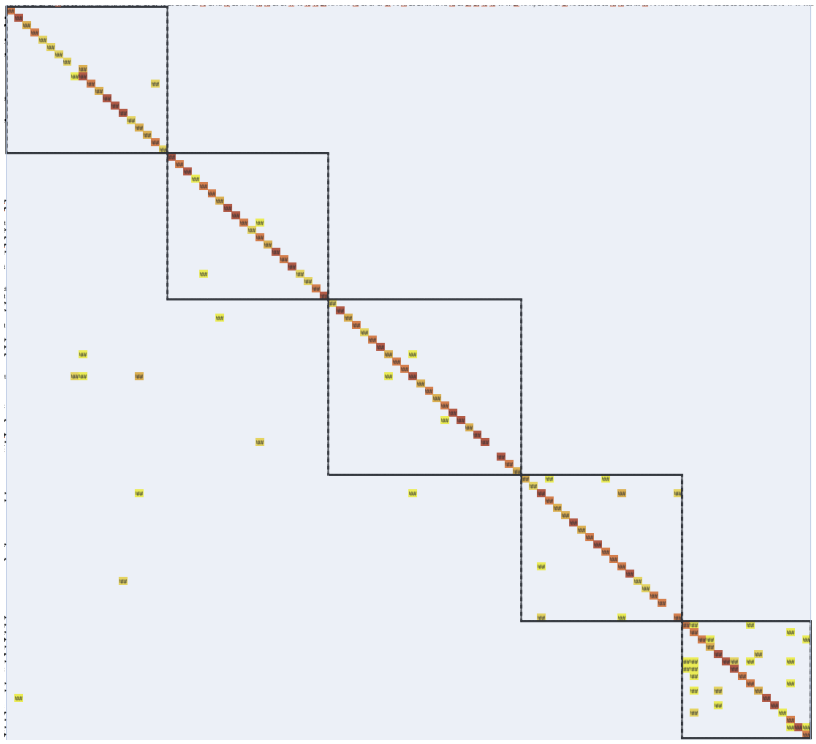


Figure 5.4: Similarity within categories using TF·IDF at Level 1 Pipeline 1. Squares correspond to categories, and the darker the points, the higher the similarity. Dark points on the diagonal are correct matches. Most of the secondary dark points are confined in a square (a single category).

Table 5.3: Similarity between entities on the Original set using TF·IDF. The numbers represent precision (P), recall (R) and F-measure (F) for both extraction methods.

Original set	**Hungarian**						**Greedy**					
	level 1			**level 2**			**level 1**			**level 2**		
	P	F	R	P	F	R	P	F	R	P	F	R
pipeline 1	1	1	1	0.94	0.94	0.94	1	1	1	0.86	0.86	0.86
pipeline 2	1	**1**	1	0.94	0.94	0.94	1	1	1	0.86	0.86	0.86
pipeline 3	1	1	1	0.94	0.94	0.94	1	1	1	0.86	0.86	0.86
pipeline 4	1	1	1	0.96	**0.96**	0.96	1	1	1	0.88	**0.88**	0.88

Table 5.4: Similarity between entities on the Original set + noise using TF·IDF. The numbers represent precision (P), recall (R) and F-measure (F) for both extraction methods.

Original set + noise	**Hungarian**						**Greedy**					
	level 1			**level 2**			**level 1**			**level 2**		
	P	F	R	P	F	R	P	F	R	P	F	R
pipeline 1	0.9	0.94	0.99	0.83	0.87	0.91	0.9	0.94	0.99	0.74	0.77	0.81
pipeline 2	0.9	**0.94**	0.99	0.83	0.87	0.91	0.9	**0.94**	0.99	0.74	0.77	0.81
pipeline 3	0.9	0.94	0.99	0.84	0.88	0.92	0.9	0.94	0.99	0.74	0.77	0.81
pipeline 4	0.9	0.94	0.99	0.85	**0.89**	0.93	0.9	0.94	0.99	0.77	**0.81**	0.85

Discussion

The results obtained by Hungarian and Greedy extraction methods showed that the highest similarity is on Original set at level 1. The preprocessing steps seem not to influence the results at level 1. The results decrease at level 2 for both extraction methods and the best results are obtained with pipeline 4. The results on the Original set are higher than on the Original set + noise. This is expected as the non-matching entities taken from the same categories can perturb similarity. The analysis of erroneous matches for the Original set showed that errors always involve the same entities. The analysis of erroneous matches for the Original set + noise shows that the erroneous matches occur between non-matching entities (the non-matching entities match between themselves), see Figure 5.5. This is particularly relevant for the Hungarian method. Such behaviour is positive as it shows that similarity works correctly.

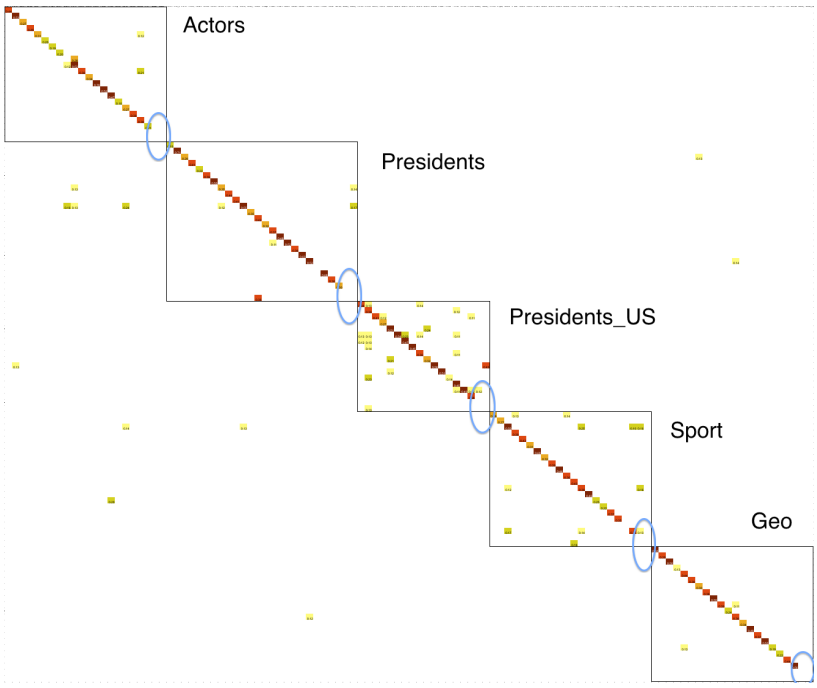


Figure 5.5: Squares correspond to categories, low similarities are between non-matching entities and are grouped in the end of each categoty.

As all results decrease at level 2, it might be possible to improve them by changing some parameter.

5.3 Experiment III: Binary Term Occurrences

This experiment is performed on Original set + noise at level 2 by changing the term weight for Binary Term Occurrences which explicitly models the absence of terms. The results are given in Table 5.5.

Table 5.5: F-measure on the Original set + noise level 2 using Binary Term Occurrences. Extraction is performed by Hungarian and Greedy methods.

Original set + noise	Hungarian	Greedy
pipeline 1	0.49	0.39
pipeline 2	0.48	0.42
pipeline 3	0.47	0.38
pipeline 4	0.62	0.46

The obtained results are significantly lower that those with have been obtained using TF·IDF. This difference demonstrates the impact of a weighting scheme in the process of selection important words for characterizing entities which are compared. The best results are obtained with pipeline 4 that may suggest that a finer filter could improve the results.

5.4 Experiment IV: Character Trigrams

This experiment is performed on Original set + noise level 2 by changing the last component of pipeline 4: instead of n-grams of terms, character trigrams are used. Moreover, cosine and Jaccard similarity measures are computed with the TF·IDF weighting scheme.

Table 5.6 shows that the best results are obtained with cosine similarity. However, modeling of documents using character trigrams has not improved the previous results. Overall, we could conclude that the standard measures such as TF·IDF and cosine similarity remain the best choice.

Table 5.6: Comparison of interlinking F-measure using cosine and jaccard similarities. Original set + noise. Level 2, pipeline 4 using character trigrams (character n-grams = 3) with TF·IDF. Extraction is performed by Greedy and Hungarian methods.

Level 2 pipeline 4	Hungarian	Greedy
cosine	0.85	0.68
jaccard	0.81	0.68

5.5 Conclusions

The results demonstrated that TF·IDF with cosine similarity and the Hungarian extraction method can identify most of the correct matches using minimum information in a resource description. The approach yielded the F-measure over 0.98 on resources representing named entities. The reported results provide evidence that machine translation can be used for finding identical resources in two different languages. It would be interesting to test if the method works at the conceptual level (resources represent thesauri concepts).

Chapter 6

Cross-lingual Linking Using Multilingual Lexicon

Abstract. In this chapter, we evaluate the BabelNet multilingual lexicon for interlinking RDF resources described in English and Chinese. Resources are represented as vectors of identifiers and similarity between resources is computed on these identifiers. The method achieves the F-measure of 0.89. The results are also compared to the translation-based method.

The previous chapter evaluated machine translation on the named entities. However, if machine translation cannot be applied due to some reason, other methods can be considered. In this chapter, we propose to use a multilingual reference resource which is one of the interlingual methods described in Chapter 3. A multilingual lexicon associates lexicalizations in different languages to identifiers which stand for concepts (senses). It can serve as a pivot language in order to make two instance representations comparable.

This chapter evaluates instance interlinking by transforming virtual documents using a multilingual lexicon, i.e., replacing terms by the corresponding entries in the lexicon. According to the general framework, labels are collected from RDF graphs and stored into virtual documents. These labels are subsequently substituted by lexicon identifiers. The identifiers are retrieved, with the help of a word sense disambiguation system, from the lexicon language version which corresponds to a language of the labels. Thus, a transformation of virtual documents of words into the virtual documents of identifiers takes place. The experiments are conducted on instances from DBpedia labeled in English and from XLore labeled in Chinese. The results are compared with the translation-based method.

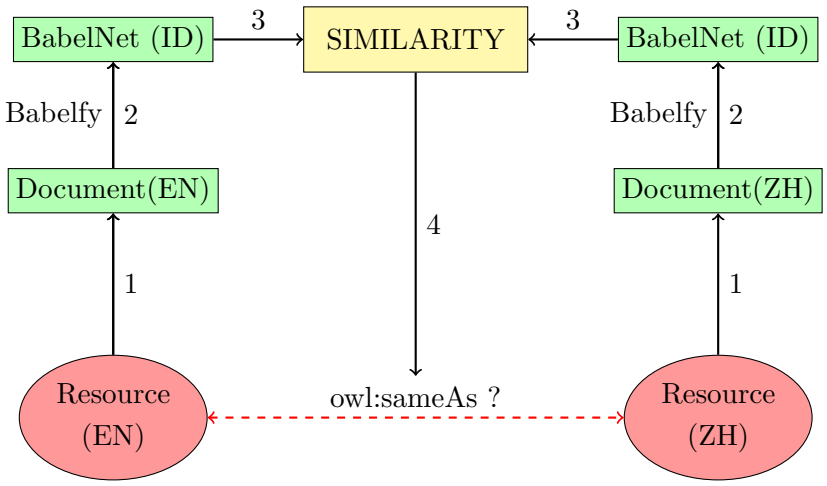


Figure 6.1: Interlinking Method Using Multilingual Lexicon. Multilingual terms are mapped to a common identifier. Similarity is computed between identifiers. Numbers correspond to the steps of the method.

Section 6.1 introduces the lexicon-based method. Instances are represented as bags of identifiers and similarity between instances is computed on identifiers. The larger overlap between bags of identifiers, the higher the chance that two representations stand for the same instance. This points to the importance of lexicon language versions to be proportional. The RDF data and experimental parameters are detailed in Section 6.2. Section 6.3 presents the results of the machine translation and multilingual lexicon-based methods. Even though the results demonstrate that machine translation outperforms its counterpart, a multilingual lexicon can be considered an appropriate intermediate for representing resources expressed in two different languages.

6.1 Lexicon-based Interlinking

The interlinking method is schematized in Figure 6.1.

In particular, the method is as follows:

1. Constructing a **Virtual Document** per resource following the procedure described in Section 4.2.

2. Replacing document terms by identifiers from a **Multilingual Lexicon** in order to project the words of each language onto the same semantic space. At this step, we represent original documents as vectors of identifiers (IDs). A corresponding identifier (ID) is retrieved for a term. An identifier stands for a sense of a term and very often there are many senses

(IDs) per term. If more that one sense exists, word sense disambiguation techniques shall be applied in order to select the best sense. The terms not found in a multilingual lexicon are discarded and we do not work with them in our experiments. To compute semantic relatedness, multilingual lexical knowledge resources can be used, e.g., BabelNet or DBnary (see Section 3.5).

3. **Computing Similarity** between documents. We use a standard term weighting scheme (TF·IDF) and apply cosine similarity. These techniques showed good performance in our previous experiments.

4. **Generating Links** between identical resources. At this stage, an algorithm extracts links on the basis of the similarity between documents. We use the Hungarian or greedy methods to extract links.

6.2 Evaluation Setup

Our goal is to evaluate how the method described above works and what parameters are important. We particularly focus on four parameters: the presence or absence of non-matching entities in a data set, the presence or absence of rdfs:label property values in a virtual document, the amount of text in a virtual document per resource and the link extraction mechanism. We evaluate the suitability of multilingual lexicon for identifying identical resources.

6.2.1 RDF Data

The experiment has been conducted on two separate RDF data sets with resources represented in English and Chinese respectively. The original data set is the same as described in Section 5.1.2, however we have enhanced it in several aspects: addition of noise and removing rdfs:label. Two datasets have been used:

- Original set: as described in Section 5.1.2;
- Original set + noise: as described in Section 5.2. Entities used as noise are entities which have been present only in one language side and have been selected from the same categories as entities from the Original set.

Each of these datasets contains virtual documents of two kinds: with an rdfs:label property value or without it. To build text collections without labels, the values of rdfs:label property are not retrieved in the English corpus. The values of http://xlore.org/property/外文名 property (meaning “Foreign name”) are not retrieved in the Chinese corpus. Thus, we have two variations of each dataset per language: Label and NoLabel.

Since we are linking named entities, an rdfs:label property value is usually a name of the entity which can be highly discriminative. By constructing a virtual document without this property value, we estimate the importance of this element in a resource description.

The average number of words in virtual documents of the Original set is 230 at level 1 and 2100 at level 2 for the English language, the numbers do not vary much when noise is added. No such statistics is available for Chinese since we do not use Chinese tokenization (it is done at lexicon-mapping step by Babelfy).

6.2.2 Experimental parameters

The parameters used for interlinking with a multilingual lexicon are presented in Table 6.1.

Table 6.1: Experimental parameters

Label 2	Data 2	VDocs 2	KB 1	Weight 1	Similarity 1	Link Extraction 2
Label NoLabel	Original set Original set + noise	level 1 level 2	BabelNet + Babelfy: EN→ID ZH→ID	TF·IDF	cosine	Greedy Hungarian

Multilingual lexicon mapping. We use BabelNet 2.5.1 which is a multilingual lexicon which connects concepts and named entities in a large network of semantic relations called synsets. Each synset represents a given meaning and contains synonyms which express that meaning in a range of different languages. Since many terms can have several synsets, we also made use of Babelfy 0.9[1] [85] in order to retrieve the best meaning per term. Babelfy had a limit of 3500 characters for input text, so we had to cut documents at level 2 only. The impact of this is that we missed additional textual information which could have been useful for similarity computation.

For illustrative purposes, an extract from a virtual document containing identifiers is given below.

[1] http://babelfy.org/

An extract from a virtual document after lexicon mapping

```
bn:00913707n
bn:00058192n
bn:01465315n
bn:00007140n
bn:00655079n
bn:00108245a
bn:00054972n
bn:00088630v
```

Machine translation. We also apply machine translation on the experimental data. We translate virtual documents using Machine Translation in order to transform documents into the same language. We use Bing Translator to translate Chinese documents into English. Once the documents are translated, we preprocess data to prepare it for similarity computation. Virtual documents are treated as "bags of words", and we use standard NLP preprocessing techniques: transform cases into lower case + tokenize + filter stop words. Once the documents are preprocessed, we apply TF·IDF and cosine similarity.

The preprocessing of virtual documents has been done using the RapidMiner toolkit with the text processing extension. The preprocessing of virtual documents undergone machine translation corresponds to Pipeline 2 described in Section 5.1.1. The preprocessing of virtual documents undergone multilingual lexicon mapping includes only tokenization according to a regular expression: ([a-z]+:). This results in suppression of "bn:" in the virtual documents.

6.3 Results

In the current evaluation, we have compared the results obtained using both methods: BabelNet and MT-based, see Table 6.2 and 6.3. We have compared the results[2] using two popular assignment algorithms: the Hungarian and greedy. The best results have been achieved by the Hungarian algorithm. Interestingly, the results of Hungarian (0.89) and Greedy (0.88) are almost the same at level 1 of Original set with Labels. However, as the testing conditions become more difficult (level 2, addition of noise, NoLabel), the Hungarian method outperforms the greedy by at least 10 points. This indicates that the global optimum is more beneficial for finding links from similarity values which can be less discriminating

[2]The lexicon-based approach has been published in [68]. However, to obtain results reported in this chapter, we modified the implementation of the Hungarian algorithm so that zero similarities are not taken into account. Due to this modification the precision improved in some cases, but the difference is negligible.

under the above mentioned conditions. The best results are obtained at level 1 on data sets with the rdfs:label property. Results at level 2 decrease for both algorithms: this is because information at level 2 becomes less discriminative and more noisy. Results are also lower when non-matching entities are added. In general, the translation approach outperformed the approach based on multilingual lexicon. This might be due to the better development of MT capability and unavailability of identifiers for some terms as well as errors in disambiguation in BabelNet. Since the terms not found in BabelNet have been discarded (as per step 2 Section 6.1), we know neither the nature of the missing terms nor the distribution of the number of missing terms per entity. If missing terms are preserved, the absence of identifiers may be compensated by translating those terms using machine translation. The results at level 2 may have been affected by the input text limit of Babelfy. The use of word sense disambiguation system (Babelfy) improved the results compared to the setting in which all identifiers are retrieved per term as reported in [66].

6.4 Conclusions

We have evaluated two approaches based on multilingual lexicon and machine translation. The best results are obtained using machine translation with an F-measure equals to 1. The F-measure of 0.89 obtained with the multilingual lexicon is slightly lower. The highest F-measure has been obtained at Level 1 on datasets with the rdfs:label property which shows that a name of a named entity is a discriminative feature in the interlinking process. Overall, both approaches seem to be promising for cross-lingual RDF data interlinking. However, the limitation would be the availability of language resources for a given pair of languages. The approach can be extended further by testing if both approaches can be complementary: errors made by one method can be corrected by the other method.

The next chapter describes the evaluation of the translation-based methods on thesauri concepts. The evaluation tests if this approach is beneficial for linking resources which are not named entities.

Table 6.2: Comparison of MT and BabelNet Methods. Similarity between Entities Using TF·IDF. The numbers represent precision (P), recall (R) and F-measure (F) for the greedy extraction method.

	Greedy	**Machine Translation**						**BabelNet**					
		level 1			**level 2**			**level 1**			**level 2**		
		P	F	R	P	F	R	P	F	R	P	F	R
Label	**Original set**	1	**1**	1	0.86	0.86	0.86	0.91	0.88	0.86	0.68	0.68	0.68
	Original set + noise	0.9	**0.94**	0.99	0.74	0.77	0.81	0.75	0.76	0.78	0.63	0.66	0.69
NoLabel	**Original set**	0.88	**0.87**	0.86	0.79	0.79	0.79	0.77	0.75	0.73	0.67	0.67	0.67
	Original set + noise	0.76	**0.79**	0.83	0.69	0.72	0.76	0.65	0.66	0.68	0.6	0.63	0.66

Table 6.3: Comparison of MT and BabelNet Methods. Similarity between Entities Using TF·IDF. The numbers represent precision (P), recall (R) and F-measure (F) for the Hungarian extraction method.

	Hungarian	**Machine Translation**						**BabelNet**					
		level 1			**level 2**			**level 1**			**level 2**		
		P	F	R	P	F	R	P	F	R	P	F	R
Label	**Original set**	1	**1**	1	0.94	0.94	0.94	0.91	0.89	0.87	0.83	0.83	0.83
	Original set + noise	0.9	**0.94**	0.99	0.83	0.87	0.91	0.76	0.77	0.79	0.7	0.73	0.77
NoLabel	**Original set**	0.94	**0.94**	0.94	0.92	0.92	0.92	0.83	0.81	0.80	0.78	0.78	0.78
	Original set + noise	0.81	**0.84**	0.88	0.78	0.82	0.86	0.74	0.76	0.78	0.65	0.68	0.71

Chapter 7

Linking Generic Entities Using Machine Translation

Abstract. In this chapter, we evaluate machine translation on terminologies expressed in different natural languages. In the evaluated experiments, we use only one pair of languages at a time, i.e., English vs. French, English vs. Chinese, German vs. French, etc. The results demonstrated that machine translation can work well independently of a dataset structure. The present evaluation shows that the translation-based method can be applied on resources which do not necessarily contain a named entity as their label.

The two previous chapters evaluated methods for interlinking named entities. The translation-based interlinking method has been applied to the encyclopedic resources in English DBpedia and Chinese XLore on which we obtained the good results presented in Chapter 5. Though this method has been initially developed for interlinking RDF instances with labels expressed in different natural languages, we consider its application to linking heterogeneous multilingual *linguistic* resources as described in Section 3.6.1. In this chapter, we consider interlinking of concepts, i.e., generic entities named with a common noun or term. Our broad goal is to evaluate techniques that make no assumption about a particular type of resources as long as these resources are published in RDF.

Section 7.1 presents the first experiment which evaluates machine translation on concepts from the TheSoz multilingual thesaurus in three languages: English, French and German. Even though the obtained results are high, it might be due to the same structure of concept descriptions as the concepts belong to the same thesaurus. To verify that machine translation results are independent of the knowledge structure, we conducted another experiment involving two different thesauri. Section 7.2 presents the second experiment which evaluates machine

translation on concepts in English and Chinese from EuroVoc and AGROVOC respectively. These two experiments demonstrate that machine translation performs well in both cases. Using the best results of these two experiments, Section 7.3 shows that similarity thresholding of the obtained links may not be very useful.

7.1 Experiment I: Linking TheSoz Concepts

7.1.1 Translation-based Interlinking Method

The interlinking approach based on machine translation technology has been already presented in Chapter 4. The interlinking method consists of five steps:

1. Constructing a **Virtual Document** in different languages per resource following the procedure of Section 4.2. At this step, we suppress all metadata information about the dataset: for example, objects of "http://purl.org/dc/terms/" property describe creators of the dataset, dates of creation and modification. The properties to remove were detected by observing the generated documents. Thus, a virtual document contains only proper lexical items, the names of the properties themselves are also omitted.

2. Translating documents using **Machine Translation** in order to transform documents into the same language.

3. Cleaning documents using **Data preprocessing** techniques. We use the following text preprocessing: Transform Cases into lower case + Tokenize + Filter stop words.

4. **Computing Similarity** between documents.

5. **Generating Links** between concepts.

An example of a virtual document at Level 1 before suppressing metadata:

```
working hours
3.3.06
```

The same virtual document at Level 1 after suppressing metadata:

```
working hours
```

An example of a virtual document at Level 2 before suppressing metadata:

```
Descriptor
Descriptors of the TheSoz
...
2011-05-06
2011-05-06
2014-08-14
0.93-en
GESIS - Leibniz-Institut für
                Sozialwissenschaften
GESIS - Leibniz Institute for the
                Social Sciences
http://www.gesis.org/das-institut/impressum/
http://www.gesis.org/en/institute/impressum/
overtime
3.3.06
working hours
agricultural working hours
management of working hours
extension of working hours
sunday work
weekend labor
Work Organization, Job Engineering, Job Satisfaction,
Industrial Safety
3.3.06
time
5.1.00
eight hour day
3.3.06
capacity-oriented variable working hours
3.3.06
flexible working hours
3.3.06
annual hours of work
3.3.03
lifetime work period
3.3.03
working week
3.3.06
```

The same virtual document at Level 2 after suppressing metadata:

```
working hours
overtime
working hours
agricultural working hours
management of working hours
extension of working hours
sunday work
weekend labor
Work Organization, Job Engineering, Job Satisfaction,
Industrial Safety
time
eight hour day
capacity-oriented variable working hours
flexible working hours
annual hours of work
lifetime work period
working week
```

7.1.2 Evaluation Setup

The main objective of this evaluation is to assess the performance of the interlinking method on resources which may not contain Named Entities as their labels.

In this section, we first describe the multilingual data used for experiments. Then we describe the parameters used for evaluating the approach.

Data

In order to conduct the evaluation, datasets with a set of reference links has to be used. As an alternative, we used one dataset with labels of the same resource in different languages. In this case, several datasets are generated according to the language of the labels and comparison is performed between these newly created datasets.

As a source of a multilingual terminological corpus, we use a multilingual thesaurus for the Social Sciences - TheSoz 0.93 in English, German and French languages mentioned in Section 3.6.1. Table 7.1 shows information about the concepts in the thesaurus. There are 8223 concepts in total for each language. 12 of them have no English label, and 6 concepts do not have French label. There are 8206 common concepts with a corresponding language label.

Table 7.1: Representation of concepts in each language version of the TheSoz

TheSoz	EN	DE	FR
total number of concepts	8223	8223	8223
concepts without label	12	0	6
number of common concepts	8206	8206	8206

In order to provide a reference alignment, we split the thesaurus into three language specific datasets which contain the same concepts with a label in a respective language. Since the same URI identifies a given concept in each language, we could compare the obtained links against the reference. The dataset consists of 223,574 triples in each language version. In the experiments, we use the 8206 concepts shared by all three languages.

Evaluated Configuration

The parameters evaluated are presented in Figure 7.1.

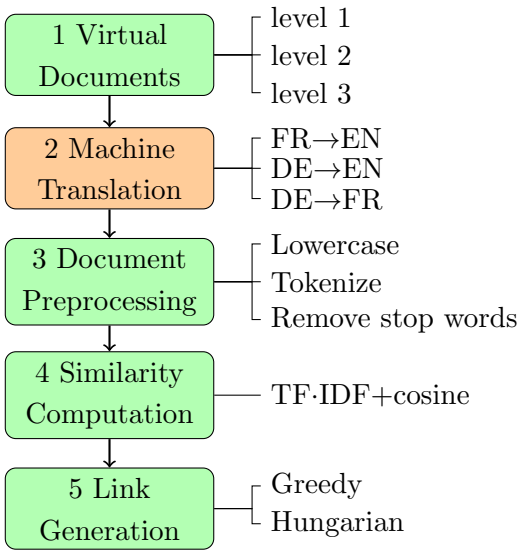


Figure 7.1: Experimental parameters.

Virtual Documents. We constructed virtual documents for Level 1 and Level 2 for the three language pairs. After the results were obtained, we decided to build virtual documents at Level 3 for the best language pair in order to see whether a larger context affects the results.

French and German translation to English. Once we collected virtual documents from the English and French/German data sets, we needed to make these documents comparable. For our experiment, we used the statistical translation system Bing Translator to translate French and German virtual documents into the English language. Thus, if we compare French virtual documents with the German ones, English is a pivot language.

German translation to French. In order to verify that the way the virtual documents are translated can affect the results, we also translate German into French, and compare the translated documents against the original French dataset. In this case, the translation is done directly from the source language (DE) into the target one (FR).

Data Preprocessing and Similarity Computation. RapidMiner 5.3.013 with the text processing extension was used for document preprocessing. Each data preprocessing step corresponds to a particular operator in RapidMiner. The following configurations were used:

- Tokenize: mode: non-letters;
- Filter Stopwords (English, French): built-in stopword lists;
- The TF·IDF weighting scheme was used in all settings;
- For computing similarity, we were using Data to Similarity Data operator with cosine similarity.

Link Generation. The output of the similarity computation is a matrix of similarity values between compared entity pairs. We use the Hungarian and greedy algorithms to extract the match assignments. All null similarities were not considered during match extraction.

Randomly removed concepts. The original 8206 concepts common to three language-specific datasets are in a one-to-one relationship with each other. We conducted an additional experiment in order to see how the similarity behaves if concepts in one dataset do not appear in the other one. This experiment has been done on the language pair which showed the highest results using the evaluated configuration described in 7.1.2: EN-DE language pair. We randomly suppressed 40% of concepts from both datasets and only 60% of the concepts has been preserved. Thus, out of 8206 original concepts, only 4943 concepts took part in the experiment. 2995 concepts constituted reference links.

Protocol

The evaluation was carried out according to the following protocol:

- Provide the two sets of resources;
- Run the method and collect the links;
- Evaluate links against the reference links through precision, recall and F-measure.

7.1.3 Results

The results where French and German virtual documents have been translated into English and compared against the original English data are provided in Figure 7.2. The results of comparison against French original data where German virtual documents have been translated into French are presented in Figure 7.3. Finally, Figure 7.4 shows the results of the additional experiment with randomly removed concepts. Each subfigure shows results for a particular language pair using both link extraction algorithms, we compute the F-measure for each setting and present it on the y-axis.

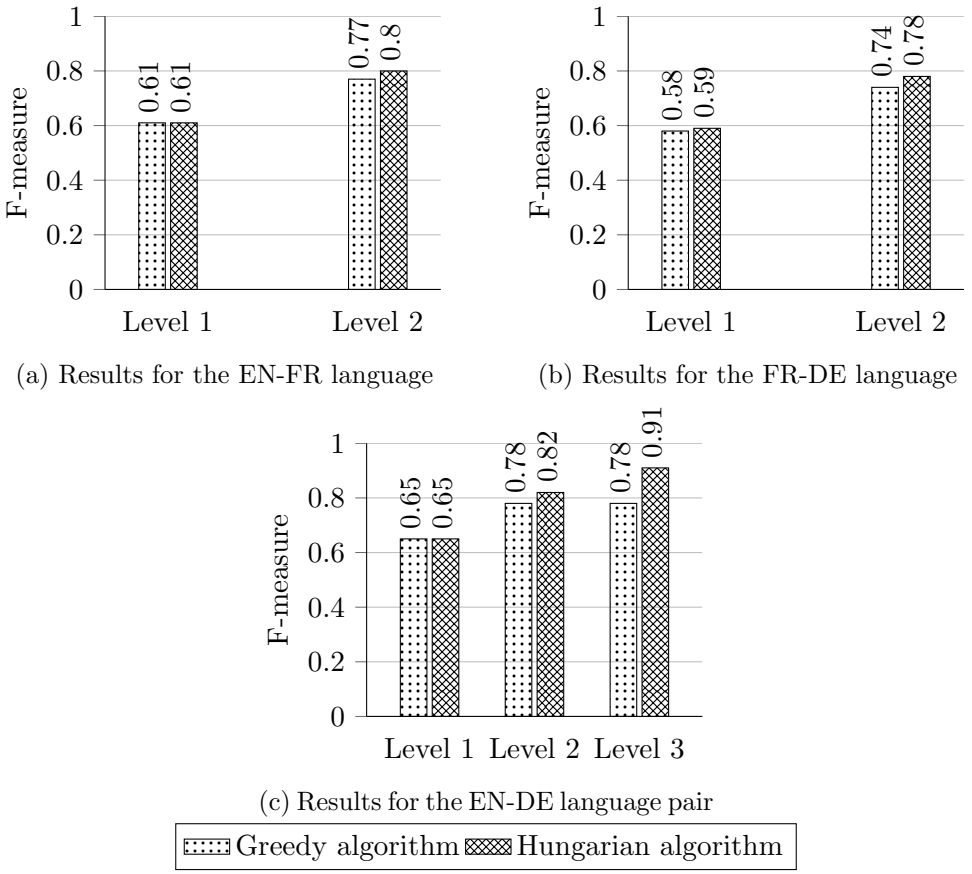
Figure 7.2, Figure 7.3 and Figure 7.4 demonstrate that the F-measure grows with level. The best F-measure of 0.91 was found at Level 3 which is an improvement of 26 percentage points compared to Level 1.

The results using English as pivot language are better than direct translation between German and French. The results where French and German virtual documents have been translated into English and compared against the original English data are provided in Figure 7.2. The results of comparison against French original data where German virtual documents have been translated into French are presented in Figure 7.3.

Figure 7.4 shows the results of the additional experiment with randomly removed concepts. The results show that the accuracy decreases when the overlap between thesauri decreases. The best matches are obtained at Level 2 and 3 with F-measure of 0.59 for the Hungarian method.

Concerning the link extraction methods, both link extraction algorithms obtained relatively similar results at Level 1. The Hungarian algorithm outperformed the greedy one at Level 2 and Level 3 and showed an increase of F-measure.

In the present experiment, the obtained results are different from results obtained with Named Entities. In previous experiments [67], the cross-lingual interlinking has been done between resources representing Named Entities, and the method could identify most of the correct matches with the F-measure over 0.95 at Level 1.



(a) Results for the EN-FR language

(b) Results for the FR-DE language

(c) Results for the EN-DE language pair

Figure 7.2: French and German languages are translated into English and compared against the English original data. For FR-DE pair, English is a pivot language. Results for Level 1, Level 2 and Level 3 using TF·IDF.

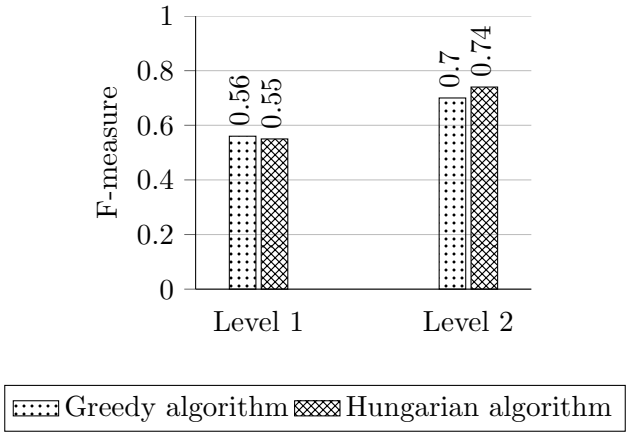


Figure 7.3: Results for the FR-DE language pair. German language is translated into French, comparison done against French original data.

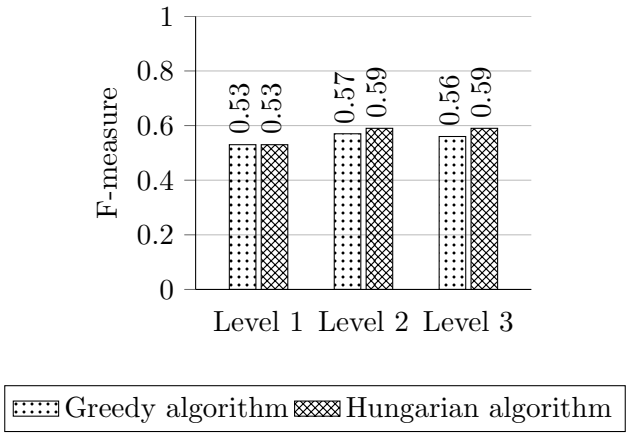


Figure 7.4: Results for the EN-DE language pair. 40% of the concepts have been randomly removed from both datasets.

The quantity of information in virtual documents can influence the output of machine translation. Level 1 often contains a single word or a short phrase. If machine translation is not exact at Level 1, the mismatch is possible. That is why it is important to extend the context of a term by proceeding to further levels.

The best results are obtained for the English-German language pair (Figure 7.2). The worst results relate to the French-German language pair when the German language has been directly translated into French (Figure 7.3).

The results of the experiment with randomly removed concepts (Figure 7.4) show again that the similarity between entities grows as the level increases: precision has been relatively the same across all levels, and we observed an increase of recall by at least 10 percentage points from level 1 to further levels. Though the results are lower, the correct matches have got the highest similarity values even when resources are not in a one-to-one relationship.

The conducted evaluation showed a different performance of the interlinking method when tested on the resources represented by generic terms (a concept label is usually a common noun or a term in a thesaurus). Thus, it seems that it is more difficult to interlink concepts of a thesauri rather than resources corresponding to named entities.

Error Analysis

We analyzed the errors occurring in the EN-DE language pair (according to the results in Figure 7.2) which showed the highest results. A false positive (FP) link is an extracted link which is not in a reference. A false negative (FN) link is a link which is in a reference but was not extracted. We specifically test if

FP and FN decrease monotonically across levels. We test it on the generated links as well as on the entities which appear in these links. We address several questions:

Q1: Do we retrieve less errors as level increases? The results are presented in Figure 7.5. We observe that less incorrect links are retrieved as level increases, in particular this observation is true for the Hungarian method. The greedy method does not show the same behavior.

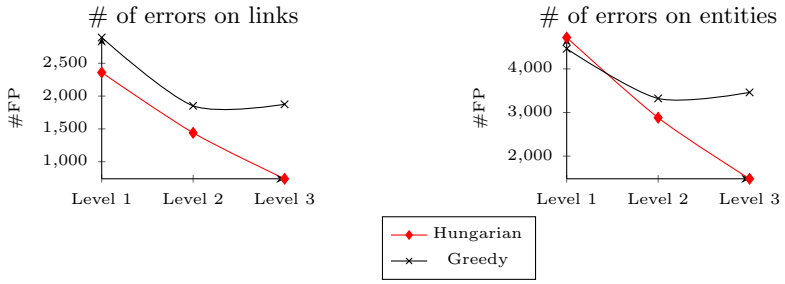


Figure 7.5: The number of FP across levels for both link extraction methods.

Q2: Do we miss less correct links as level increases? The results are presented in Figure 7.6. We observe that the number of correct links which are missed decreases as level increases. This observation is true for both link extraction methods.

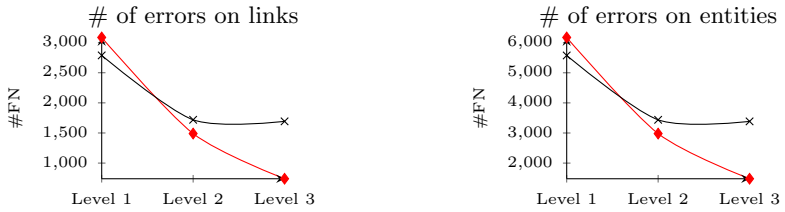


Figure 7.6: The number of FN across levels for both link extraction algorithms.

Q3-4: Do the link extraction algorithms make the same errors across levels? Are the missed links the same across levels? To that extent we measured:

- the ratio of new False Positives (FP) introduced when level n increases:

$$\frac{|\ FP_{n+1} \setminus FP_n\ |}{|\ FP_{n+1}\ |};$$

- the ratio of new False Negatives (FN) when level n increases:

$$\frac{|\ FN_{n+1} \setminus FN_n\ |}{|\ FN_{n+1}\ |}.$$

These two ratios have been computed both on links (see Table 7.2) and entities (see Table 7.3). Computing these two ratios on links allow to see if wrong and missed links are the same when level increases. Nevertheless, it could happen that the wrong or missed links are not the same but are made on the same entities. To that extent, we also computed them on entities which appear in the found links (only unique occurrence of an entity is taken into account (duplicates are removed)).

Table 7.2: Wrong and missed links introduced across levels.

Greedy	L1 → L2	L2 → L3
FP	0.74	0.92
FN	0.08	0.42

Hungarian	L1 → L2	L2 → L3
FP	0.79	0.77
FN	0.05	0.11

Table 7.3: Wrong and missed entities introduced across levels.

Greedy	L1 → L2	L2 → L3
FP	0.34	0.44
FN	0.08	0.42

Hungarian	L1→ L2	L2→ L3
FP	0.30	0.16
FN	0.05	0.11

We observe that, when level increases, among the wrong links, many are not present at the previous level. But, these errors are in the majority made on the same entities even if there are, in average, around 30% of new entities in the introduced wrong links. A further analysis shows that more than 80% of entities that appear in introduced wrong links at level 2 were in the missed links at level 1. From level 2 to 3, this drops to 28% for Hungarian and 11% for greedy.

In terms of missed links, we can see that they tend to be included in the set of links missed at lower level. Once again, Hungarian performs better than greedy.

Discussion

This experiment showed a different performance of the interlinking method when tested on the resources represented by generic terms (a concept label is usually a common noun or a term in a thesaurus). Thus, it seems that it is more difficult to interlink concepts of a thesauri rather than resources corresponding to named entities. The finding suggests that the interlinking strategy (including the automatic selection of levels) may depend on the type of entities to be interlinked. Chapter 8 describes a hypothesis that comparison of entities belonging to different types can be done at different information levels.

7.2 Experiment II: Linking EuroVoc-AGROVOC Concepts

The translation-based interlinking method described in Section 7.1.1 has been evaluated on concepts from EuroVoc and AGROVOC thesauri.

Data

We use multilingual thesauri from multidisciplinary and agricultural domains: EuroVoc and AGROVOC. We extracted entities from the existing reference alignment (1318 entities linked by "skos:exactMatch" property). We suppressed duplicate concepts from EuroVoc and their corresponding concepts from AGROVOC. In the experiments, we use the 1300 concepts in English from EuroVoc and in Chinese from AGROVOC. The reference contains 1300 links in which concepts are in one-to-one correspondence. The evaluated parameters remained the same as described in Figure 7.1 except that the Chinese labels have been translated into English.

7.2.1 Results

The results where Chinese virtual documents have been translated into English and compared against the original English data are provided in Figure 7.7. The main difference with the TheSoz results is that F-measure drops as levels grow. The best F-measures of 0.81 and 0.80 at Level 1 were obtained by both link extraction algorithms[1]. The results at Level 3 dropped significantly (by 20 percentage points) for both algorithms. The decrease of the results at Level 3 can be due to the difference in knowledge organization of each thesaurus.

7.3 Comparison of Results According to a Threshold

The results of both link extraction algorithms are evaluated according to a threshold. Figures 7.8 and 7.9 present the best results of the TheSoz concept linking, i.e., for the English-German language pair according to the results in Figure 7.2. Figures 7.10 and 7.11 present the results of the EuroVoc-AGROVOC concept linking. The threshold corresponds to a similarity value for which extracted links were evaluated. The purpose of this evaluation was to observe if the results change drastically after a certain threshold. We could observe that the F-measure decreases in all cases because recall decreases faster than precision increases. Overall, the correct matches are distributed across a wide range of

[1] An F-measure of 0.82 is reported in [30]. However, the number of reference links reported is different.

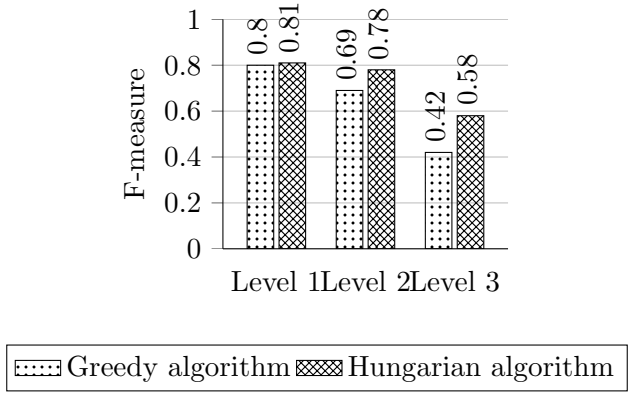


Figure 7.7: Results on concepts from EuroVoc-AGROVOC on the EN-ZH language pair.

similarity values, so establishing the threshold above 0 may not provide the best cutoff.

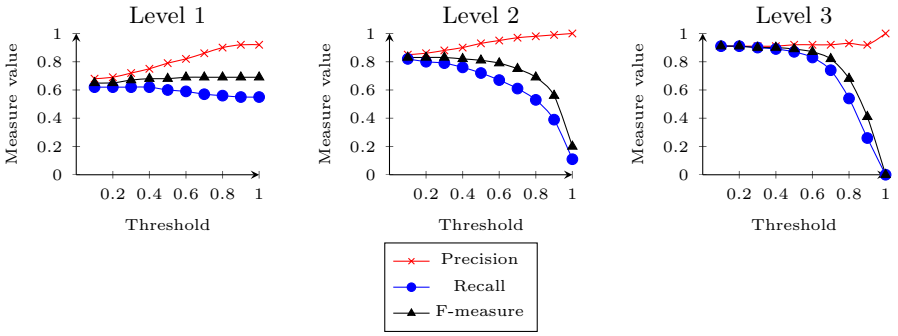


Figure 7.8: Hungarian results for TheSoz: the EN-DE language pair.

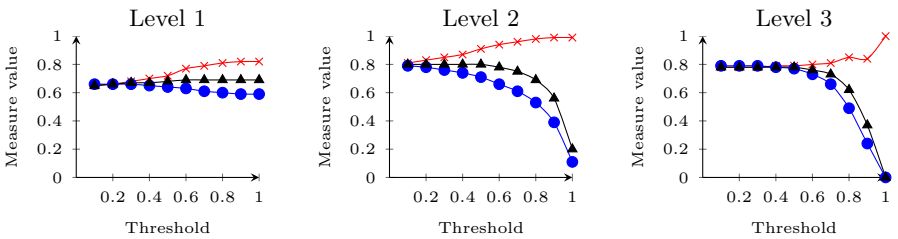


Figure 7.9: Greedy results for TheSoz: the EN-DE language pair.

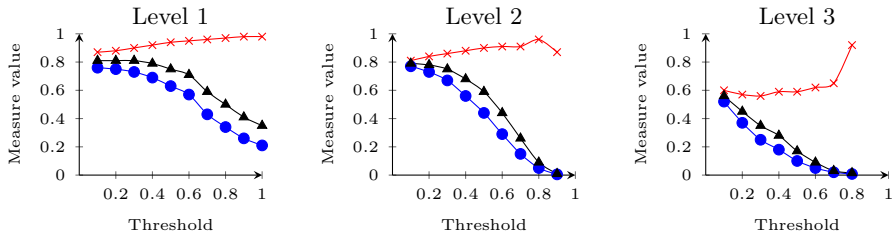


Figure 7.10: Hungarian results for EuroVoc-AGROVOC: the EN-ZH language pair.

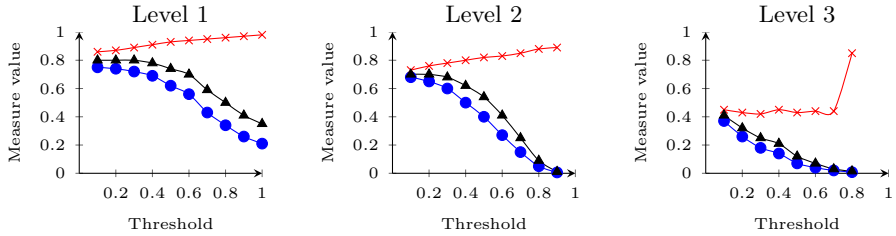


Figure 7.11: Greedy results for EuroVoc-AGROVOC: the EN-ZH language pair.

7.4 Conclusions

This chapter evaluated machine translation on interlinking terminology expressed in different natural languages. We observed the impact of the quantity of textual information in resource description by collecting information from further removed neighboring nodes. We evaluated the approach on 8206 thesaurus concepts in English, French and German languages from the social science domain. We compared the generated links of the Hungarian and greedy assignment algorithms. In our previous evaluation performed on English-Chinese Named Entities from RDF encyclopedias (DBpedia and XLore), the highest results have been achieved at Level 1 with precision over 0.98. In contrast to those results, the best results have been obtained at Levels 2 and 3. The highest result with an F-measure of 0.91 has been obtained at Level 3 for the EN-DE language pair. The best correspondences have been extracted by the Hungarian algorithm. The best experimental results on EuroVoc-AGROVOC concepts achieved the F-measure of 0.81. These results obtained using machine translation on labels from one language (Chinese) are comparable to the results obtained using multiple labels of the concepts.

The results of both experiments demonstrate that machine translation can work well independently of a dataset structure. The present evaluation shows that the translation-based method can be applied on resources which do not necessarily contain a named entity as their label, though it is harder to find a

correct correspondence in this case. Overall, the proposed method is a practical way to interlinking RDF linguistic resources since it does not depend on a rich multilingual representation of concepts.

There are several parameters for further investigation:

- Use of other machine translation engines;
- Use of external lexical resources for language mediation.

The next chapter contains several perspectives on cross-lingual data interlinking. Its first section describes the hypothesis mentioned in Section 7.1.3. Then, we propose to modify the construction of virtual documents and to use other cross-lingual techniques. Finally, three ways of combining machine translation and lexicon-based methods are discussed.

Chapter 8

Perspectives

Abstract. In this chapter, we propose several directions of research on evaluation of cross-lingual techniques for data interlinking. We propose a hypothesis about linking two kinds of resources as well as three ways of combining machine translation with multilingual lexicons. The hypothesis suggests that collecting textual information for resource interlinking can depend on the nature of resources to be interlinked. For named entities, the closest neighborhood can be sufficient to identify identical resources. For generic concepts, further removed neighbors may be necessary. Combination of machine translation with multilingual lexicons may help to compensate for the shortcomings of each method.

This thesis provides insightful results about the usability of natural language resources for cross-lingual data interlinking. It also raises some questions and some perspectives. We consider these here and suggest experiments for testing them using or extending our experimental framework.

The previous chapters showed that interlinking methods behave differently on different kinds of RDF resources. Section 8.1 proposes an experiment to test the hypothesis that, when generating cross-lingual links between different language descriptions of resources, graph traversal should be limited for named entities and more extensive for generic terms. This hypothesis arose from the obtained results presented in the previous experiments.

Section 8.2 and Section 8.3 propose different NLP techniques that were not tested so far.

Section 8.2 considers a modification of the first component of the proposed framework (see Chapter 4). It proposes to create coherent texts from RDF triples instead of collecting separate literals.

Section 8.3 considers a modification of the second component of the proposed

framework: other cross-lingual techniques can be tested further regarding their utility for data interlinking.

Section 8.4 considers influence of type of input. It suggests that cross-lingual interlinking can be performed on domain specific knowledge.

The experimental results indicated that machine translation outperformed the lexicon-based approach though it is not obvious how each of the methods could be complementary to the other. Section 8.5 combines tested approaches. It shows how machine-translation and lexicon-based interlinking methods can be joined. Three scenarios of combining these methods are described.

8.1 Testing if Neighbors in RDF Graphs Identify Differently Named Entities and Generic Terms

Resources can be any kind of entities: in particular, Named Entities, e.g., actors, presidents, geographical places or generic terms (common nouns), e.g., altruism, labor, cognition.

In previous experiments described in Chapter 5, cross-lingual interlinking has been performed between encyclopedic resources representing Named Entities, and the method could identify most of the correct matches with F-measure over 0.95 at level 1. In subsequent experiments shown in Chapter 7, cross-lingual interlinking has been conducted between thesaurus concepts representing generic entities named as common nouns or terms. The best F-measure of 0.91 was found at level 3 which is an improvement by 26 points compared to level 1. These results show an opposite tendency. This observed phenomenon requires further investigation. We hypothesize that the number of levels for resource interlinking depends on the kind of resources to be interlinked.

This section presents the design of an experiment for testing this hypothesis: Named Entities are better interlinked at limited depth while concepts are better matched at greater depth. Considering that the previous experiments have been carried out on data sets of different languages and different characteristics, a new experiment for testing the hypothesis is necessary.

The hypothesis to be tested is described in the next section. The strategy adopted for interlinking resources of different kinds is described in Section 8.1.2.

8.1.1 Hypothesis

The hypothesis to test is as follows:

If RDF resources represent Named Entities, interlinking gives better results at Level 1. If RDF resources represent generic concepts, interlinking gives better results at Level 2 and higher.

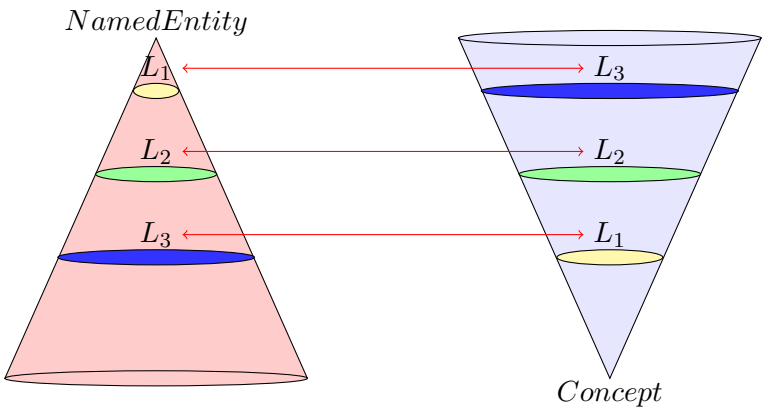


Figure 8.1: Interlinking RDF data of different kinds. For Named Entities, information becomes more general at Level 3. For concepts, the most general information is at Level 1. The more general information, the less discriminative it becomes.

As shown in Figure 4.3, an RDF graph for a particular resource can be decomposed into n levels (level 1, 2 and so on). We suppose that comparison of resources belonging to different kinds of entities can be done at different levels. If data is about Named Entities, then it is sufficient to collect information from the resource's closest literals; the further we traverse the graph, the more noise is introduced into the description of the resource (many resources will have similar information and it is harder to find an equivalent entity). Hence, increasing n should increase recall and decrease precision. If data is about generic concepts (as in thesauri), then the further we traverse the graph, the more discriminant information becomes, and it is easier to find equivalent entities. The intuition for this is that the abstract concepts will have to be explained in concrete terms at some point in a graph. These concrete descriptions of abstract terms are more discriminant. In other words, as the traversal distance grows, the neighborhood of named entities will include more generic terms whereas the neighborhood of generic terms will include more specific terms. Thus, these two types of data (named entities and concepts) can be in opposition to each other as depicted in Figure 8.1.

8.1.2 Method

Following the framework discussed in Chapter 4, the method consists of the steps described in Figure 8.2.

More precisely, the method is as follows:

1. Constructing a **Virtual Document** per resource as described in Sec-

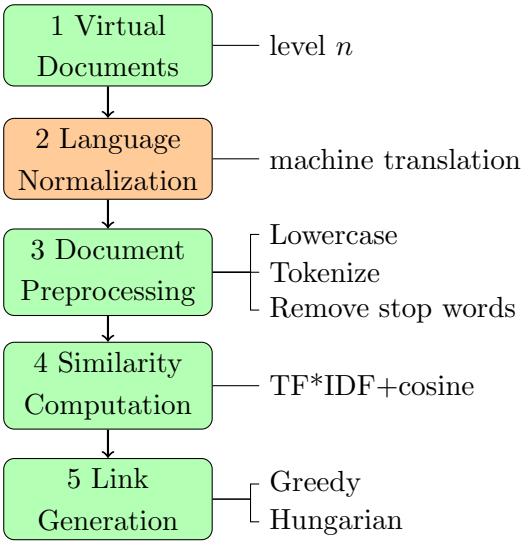


Figure 8.2: Experimental setting.

tion 4.2. At this step, all information which is not very descriptive, i.e., names of the ontology classes to which such resources are instances of (objects of rdf:type property) is suppressed. Thus, a virtual document contains only proper lexical items (literals).

2. Translating documents using **Machine Translation** in order to transform documents into the same language. At this step, virtual documents in one language can be translated into the other language and vice versa or both languages can be translated into some pivot language. Google Translate can be used to translate a source language into a target language.
3. Cleaning documents using **Data preprocessing** techniques such as tokenization, stop-word removal. The following text preprocessing is applied: transform cases into lower case + tokenize + filter stop words.
4. **Computing Similarity** between documents. The standard term weighting scheme (TF·IDF) and cosine similarity are used. These are the classical techniques for finding similar documents, moreover, they showed good performance in our previous experiments. The output of this step is a set of similarity values between pairs of virtual documents.
5. The goal of **Generating Links** is to identify a set of correspondences between concepts. At this stage, an algorithm extracts links on the basis of the similarity between documents. The Hungarian method which maximizes the sum of similarities is used to extract correspondences.

8.2 Natural Language Generation for Virtual Document Construction

Machine translation as well as word sense disambiguation systems generally work better on coherent natural language texts instead of words assembled together accidentally. Thus, it would be useful to test the influence of full-fledged sentences on the quality of generated links. Natural language generation from RDF representations aims at generating human-readable texts from RDF descriptions [40, 80, 127]. RDF representations containing linguistic information in machine-readable mark-up (e.g., class names, property names) are exploited for generating sentences describing RDF resources. In our proposed framework, instead of collecting available literals and storing them in virtual documents, natural language generation may be applied on triples in order to generate virtual documents made of sentences.

8.3 Evaluating other Cross-lingual Techniques for Language Normalization

Evaluation of other cross-lingual techniques for data interlinking is another interesting direction. For instance, as discussed in Chapter 3, the Explicit Semantic Analysis could be used as an interlingual method. It would project two resource descriptions into a common space of Wikipedia articles, and, due to the presence of cross-language links between the articles, resources can be compared. This method is sensitive to the information coverage in a particular language. For example, Wikipedia in Chinese is smaller in size compared to Wikipedia in English. As a consequence, insufficiency of knowledge in one of the language versions may impact the results.

8.4 Application of Cross-lingual Techniques to Domain Specific Knowledge

So far, machine translation and mapping to multilingual resources have been applied on data of general interest (the encyclopedic resources). However, the proposed methods could be applicable to domain specific data. We did not investigate this aspect though machine translation has been tested on AGROVOC, EuroVoc and TheSoz thesauri which are from narrow domains. The major restrictive component is a language-specific component. Statistical machine translation can be trained on parallel corpora from a specific domain given the availability of such corpora. Moreover, specialized dictionaries can be useful. Such

dictionaries can be plugged into a machine translation engine for terminology recognition. The same difficulty will be faced by multilingual resources harvested from resources belonging to general domains. Hence, their application might be less effective. Overall, cross-lingual techniques can be useful for linking technical vocabularies. In this case, the availability of domain specific language resources is a necessary prerequisite.

8.5 Combining MT and Lexicon for Cross-lingual RDF Data Interlinking

In previous chapters, we described experiments on cross-lingual RDF data interlinking using machine translation and lexicon-based methods separately. The machine translation and lexicon-based approaches showed good results in these experiments. Even though the machine translation approach showed better results, we consider that both of these methods can be complementary and can be applied jointly to cross-lingual data interlinking task.

In this section, we discuss the combination of machine translation and multilingual lexicons in order to find identical resources.

Three scenarios in which machine translation and a multilingual lexicon work together are presented below.

Scenario 1

The interlinking process is schematized in Figure 8.3.

The method consists of the following steps:

1. Constructing a **Virtual Document** in different languages per resource.

2. Translating documents using **Machine Translation** in order to transform documents into the same language. Virtual documents in source languages are translated into a target language or both source languages can be translated into some pivot language. In addition to translation, the virtual document terms can be replaced by identifiers from a **Multilingual Lexicon**. Both resource representations (language words and lexicon identifiers) will be merged.

3. Cleaning documents using **Data preprocessing** techniques. The following text preprocessing can be used: Transform Cases into lower case + Tokenize + Filter stop words. This preprocessing is done only on virtual documents in target language (original and translated).

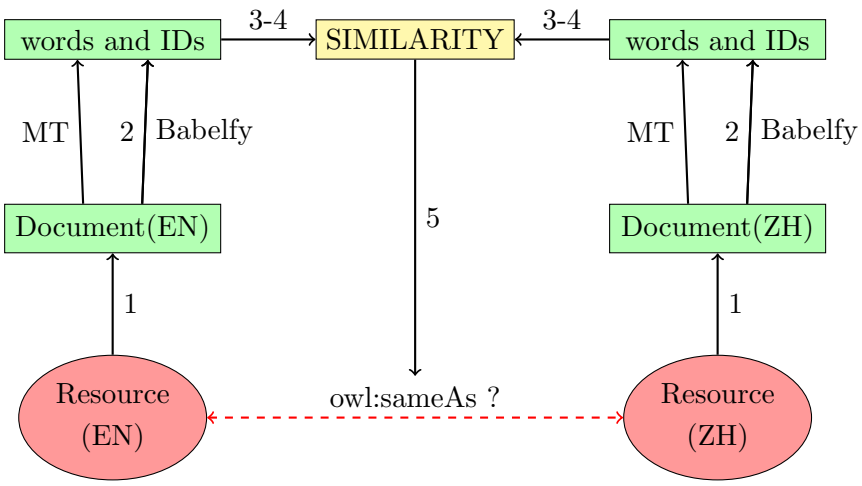


Figure 8.3: Scenario 1. Interlinking Method Combining MT and Multilingual Lexicon. MT and Lexicon mapping are applied to data in parallel. Similarity computation is performed on both natural language descriptors and lexicon's identifiers. Numbers correspond to the steps of the method.

4. **Computing Similarity** between documents using term weights and applying similarity methods, for example, the cosine similarity. The output of this step is a set of similarity values between pairs of virtual documents.

5. **Generating Links** between concepts. At this stage, an algorithm extracts links on the basis of the similarity between documents.

Scenario 2

In the second scenario, the method remains the same except the step 2: Document terms are replaced by identifiers from a Multilingual Lexicon in order to project the words of each language onto the same space. Original documents are represented as vectors of identifiers (IDs). However, some words may not be found in a lexicon. These missing terms are collected and translated using machine translation. After being translated, they are injected into the documents containing identifiers.

The difference between Scenario 1 and Scenario 2 is the the step 2. In scenario 2, machine translation is applied only on terms for which there are no lexicon identifiers; whereas, in scenario 1, machine translation and lexicon mapping are applied on the original virtual documents in parallel.

Scenario 3

In Scenario 3, similarity between resources is computed separately using machine translation and lexicon mapping (virtual document words and lexicon identifiers are never merged). Scenario 3 allows to experiment with the output of step 4. The output of Similarity Computation are two sets of similarity values: one contains similarity values of machine translation method, the other contains the results of lexicon mapping. These similarities can be combined by an aggregation function such as the average or the maximum similarity for each pair of resources. Taking the average of similarities can compensate low results produced by one of the methods. Taking the maximum value may ensure the higher precision of results. This approach remains different than that which consists of generating two sets of links by the two independent methods and merging them by either the Hungarian method or a disambiguation method.

8.6 Conclusions

Interlinking cross-lingual resources can improve discovery of facts about the same resource described in different languages. The availability of different types of RDF data and various cross-lingual techniques opens new directions of research in cross-lingual data interlinking. We observed an interesting phenomenon in previous experiments applying our interlinking method. We assume that it is due to the nature of resources to be linked (Named Entities vs. generic terms). We proposed an experiment for evaluating this hypothesis. The proposed experiment may reveal valuable insights in how similarity is affected by the nature of the resources involved.

We outlined three scenarios of how machine translation and lexicon-based methods can be combined. The proposed scenarios can be applied on any RDF dataset containing textual information in different languages.

Chapter 9

Conclusion

With the growing amount of heterogeneous data on the web, it is important to make these data machine processable. The Semantic Web provides technologies such as the Resource Description Framework (RDF) for representing data on the web. However, RDF data can be expressed with labels in different languages. Hence, data interlinking requires specific approaches to tackle multilingualism.

Cross-lingual data interlinking consists in discovering links between identical resources across RDF data sets in different languages. The use of different languages makes the comparison of these resources challenging.

Previously conducted evaluations discussed in the state of the art are limited in that the evaluated techniques have been applied independently. This does not allow to determine their benefits, fragilities and, eventually, to compare them.

This study investigated the benefit of several techniques for data interlinking across languages. For that purpose, a general framework is proposed which allows for comparing these techniques in a unified manner. The efficiency of the techniques is evaluated by conducted experiments.

The obtained results show that the transformation of RDF resources into virtual documents and the application of machine translation and multilingual lexicons are beneficial techniques for interlinking data in different languages. Virtual documents are built by collecting symbolic information (datatype property values). The resource representation as a virtual document allows to accumulate textual description of the resource and to build a context which will help to discriminate a particular resource against other resources. In the environment in which data publishers may use their own natural language to publish RDF data, it proved useful to apply machine translation and multilingual lexicons in order to render comparable the descriptions of RDF resources. The proposed approaches have been evaluated on RDF resources coming from encyclopedias such as DBpedia (English), XLore (Chinese) and thesauri (English, French, German, Chinese).

The major findings are:

- machine translation approaches showed better results compared to the lexicon mapping;
- the use of Babelfy could improve the results by 20 points compared to a setting in which all synsets are used for similarity computation;
- TF·IDF is the best term weight combined with cosine;
- n-grams can be useful in a complex setting (noisy entities and absence of entity's name);
- level 1 works best for named entities;
- levels 2 and 3 work best for thesauri concepts;
- the Hungarian method is best for link extraction.

Table 9.1 presents the best results on cross-lingual data interlinking discussed in this thesis. The difference in the results can be due to the datasets on which evaluations have been conducted. MultiFarm consists of ontologies in different languages which contain little textual information in concept descriptions. Overall, the discrepancy of the results justifies a posteriori the need for controlled evaluation of these techniques. It would be useful to conduct further investigation in reason of this discrepancy.

Table 9.1: Best Results on Cross-lingual Data Interlinking.

experiment/system	F-measure	language pair	multilingual technique
OAEI 2013	0.17	MultiFarm	CL-ESA
OAEI 2014	0.54	MultiFarm	Bing translator
OAEI 2015	0.51	MultiFarm	Bing translator
OAEI 2015	0.14	MultiFarm	BabelNet
[30]	0.82	multilingual	multilingual labels
IM@OAEI 2014	0.56	English-Italian	Google translator
Chapter 5	1	English-Chinese	Bing translator
Chapter 6	0.89	English-Chinese	BabelNet
Chapter 7	0.81	English-Chinese	Bing translator
Chapter 7	0.91	English-German	Bing translator

Multilingual resources (machine translation systems, dictionaries, knowledge-bases, encyclopedias) play an important role in a cross-lingual data interlinking task and are valuable tools for multilingual information processing. Linking entities in a multilingual context relies heavily on such resources. Interlinking RDF resources in different languages can help to uncover the potential of vast amounts of linked open data and to facilitate knowledge discovery across language bar-

riers. The results showed that linguistic resources provide enough quality to interlink data.

The presented results are promising for further exploration of the cross-language techniques. This thesis laid the ground for future systematic analyses of these techniques. Chapter 8 discussed several perspectives on cross-lingual data interlinking, in particular:

- Testing the hypothesis that closest neighborhood can be sufficient to identify identical named entities, while further removed neighbors may be necessary for generic terms;
- Use of natural language generation for virtual document construction;
- Evaluation of other cross-lingual techniques for data interlinking;
- Application of cross-lingual techniques on RDF data from specialized fields;
- Combination of machine translation with external multilingual resources in order to obtain a synergistic effect of both methods.

Additionally, studies can be extended to a larger scale due to the availability of RDF data sources. However, each experiment described in this thesis is dependent on the availability of data sets with reference links. The absence of stable benchmarks for cross-lingual data interlinking presents an impediment to evaluations. Thus, there is a need for comprehensive benchmarks in cross-lingual data interlinking, covering various languages and various types of entities.

Finally, the state-of-the-art approaches extensively use language-specific resources. Such resources can be limited for a particular language pair. Thus, it is necessary to explore language-independent methods for multilingual data processing.

Bibliography

[1] Riccardo Albertoni, Monica De Martino, Sabina Di Franco, Valentina De Santis, and Paolo Plini. EARTh: An Environmental Application Reference Thesaurus in the Linked Open Data cloud. *Semantic Web journal (SWJ)*, 5:165–171, 2014.

[2] James Allan. Topic detection and tracking. chapter Introduction to Topic Detection and Tracking, pages 1–16. Kluwer Academic Publishers, Norwell, MA, USA, 2002. ISBN 0-7923-7664-1. URL `http://dl.acm.org/citation.cfm?id=772260.772262`.

[3] Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. Towards an automatic creation of localized versions of dbpedia. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 494–509, 2013. URL `http://dx.doi.org/10.1007/978-3-642-41335-3_31`.

[4] Samur Araújo, Jan Hidders, Daniel Schwabe, and Arjen P. de Vries. SERIMI - Resource Description Similarity, RDF Instance Matching and Interlinking. *CoRR*, abs/1107.1104, 2011.

[5] Manuel Atencia, Jérôme David, and Jérôme Euzenat. Data interlinking through robust linkkey extraction. In *Proc. 21st European Conference on Artificial Intelligence (ECAI), Praha (CZ)*, pages 15–20, 2014. URL `ftp://ftp.inrialpes.fr/pub/exmo/publications/atencia2014b.pdf`.

[6] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, volume 4825, pages 722–735. Springer Berlin Heidelberg, 2007.

[7] Amit Bagga and Breck Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proc. 36th Annual Meeting of*

the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, volume 1, pages 79–85, 1998.

[8] Alberto Barrón-Cedeño, Parth Gupta, and Paolo Rosso. Methods for cross-language plagiarism detection. *J. Knowl.-Based Syst.*, 50:211–217, 2013.

[9] Biligsaikhan Batjargal, Takeo Kuyama, Fuminori Kimura, and Akira Maeda. Identifying the same records across multiple ukiyo-e image databases using textual data in different languages. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 193–196, Piscataway, NJ, USA, 2014. IEEE Press. ISBN 978-1-4799-5569-5. URL `http://dl.acm.org/citation.cfm?id=2740769.2740801`.

[10] Zohra Bellahsène, Angela Bonifati, and Erhard Rahm, editors. *Schema matching and mapping*. Data-centric systems and applications. Springer, Heidelberg, New York, 2011. ISBN 978-3-642-16517-7. URL `http://opac.inria.fr/record=b1132797`.

[11] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 37(4):573–595, 1995. ISSN 0036-1445. URL `http://dx.doi.org/10.1137/1037127`.

[12] Mikhail Bilenko, Raymond Mooney, William Cohen, Pradeep Ravikumar, and Stephen Fienberg. Adaptive Name Matching in Information Integration. *IEEE Intelligent Systems*, 18(5):16–23, September 2003. ISSN 1541-1672. URL `http://dx.doi.org/10.1109/MIS.2003.1234765`.

[13] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009. URL `http://dx.doi.org/10.4018/jswis.2009081901`.

[14] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3): 154–165, 2009.

[15] Christian Boitet. A rationale for using unl as an interlingua and more in various domains. *UNIVERSAL NETWORKING LANGUAGE: Advances in Theory and Applications*, pages 3–9, 2005.

[16] Paul Buitelaar, Key-Sun Choi, Philipp Cimiano, and Eduard H. Hovy. The Multilingual Semantic Web (Dagstuhl Seminar 12362). *Dagstuhl Reports*, 2 (9):15–94, 2013. ISSN 2192-5283. doi: http://dx.doi.org/10.4230/DagRep.2.9.15. URL `http://drops.dagstuhl.de/opus/volltexte/2013/3788`.

[17] Elena Cabrio, Philipp Cimian, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Christina Unger, and Sebastian Walter. QALD-3: Multilingual Question Answering over Linked Data. In *Working Notes for CLEF 2013 Conference*, volume 1179, Valencia, Spain, September 23-26 2013. CEUR-WS.org.

[18] Elena Cabrio, Julien Cojan, Fabien Gandon, and Amine Hallili. Querying multilingual DBpedia with QAKiS. In *Extended Semantic Web Conference (ESWC), Demo paper*, Montpellier, France, 2013.

[19] Zdenek Ceska, Michal Toman, and Karel Jezek. Multilingual plagiarism detection. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '08, pages 83–92, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85775-4. URL `http://dx.doi.org/10.1007/978-3-540-85776-1_8`.

[20] Michelle Cheatham and Pascal Hitzler. String similarity metrics for ontology alignment. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, pages 294–309, 2013. URL `http://dx.doi.org/10.1007/978-3-642-41338-4_19`.

[21] Aitao Chen and Fredric C. Gey. Experiments on cross-language and patent retrieval at ntcir-3 workshop. In *Proceedings of NTCIR-3*, 2003.

[22] Peter Chew and Ahmed Abdelali. Benefits of the 'Massively Parallel Rosetta Stone': Cross-Language Information Retrieval with over 30 Languages. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 872–879, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P07-1110`.

[23] Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, 52(3):245–275, 2011.

[24] Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Linking linguistic resources: Examples from the Open Linguistics Working Group. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing Language Data and Metadata*, pages 201–216. Springer, 2012.

[25] Peter Christen. *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data Centric Systems and Applications. Springer-Verlag, Berlin, 2012. ISBN 978-3-642-31163-5.

[26] Philipp Cimiano, Vanessa Lopez, Christina Unger, Elena Cabrio, Axel-Cyrille Ngonga Ngomo, and Sebastian Walter. Multilingual question answering over linked data (QALD-3): lab overview. In *The 4th International Conference of the CLEF Initiative, CLEF 2013*, volume 8138, pages 321–332, Valencia, Spain, September 23-26 2013. URL `http://dx.doi.org/10.1007/978-3-642-40802-1_30`.

[27] Li Ding, Joshua Shinavier, Tim Finin, and Deborah L. McGuinness. owl:sameas and linked data: An empirical study. In *Proceedings of the Second Web Science Conference*, Raleigh NC, USA, April 2010.

[28] Li Ding, Joshua Shinavier, Zhenning Shangguan, and Deborah L. McGuinness. SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl: sameAs in Linked Data. In *The 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, pages 145–160, 2010. URL `http://dx.doi.org/10.1007/978-3-642-17746-0_10`.

[29] Zlatan Dragisic, Kai Eckert, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, Stefano Montanelli, Heiko Paulheim, Dominique Ritze, Pavel Shvaiko, Alessandro Solimando, Cássia Trojahn dos Santos, Ondrej Zamazal, and Bernardo Cuenca Grau. Results of the ontology alignment evaluation initiative 2014. In *Proceedings 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*, pages 61–104, 2014. URL `http://ceur-ws.org/Vol-1317/oaei14_paper0.pdf`.

[30] Mauro Dragoni. Exploiting multilinguality for creating mappings between thesauri. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC'15, pages 382–387, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3196-8. URL `http://doi.acm.org/10.1145/2695664.2695768`.

[31] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[32] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.

[33] Ivan A. Fellegi and Alan Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[34] Alfio Ferrara, Andriy Nikolov, and François Scharffe. Data linking for the semantic web. *International Journal of Semantic Web and Information Systems*, 7(3):46–76, 2011.

[35] John Rupert Firth. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32, 1957.

[36] Bo Fu, Rob Brennan, and Declan O'Sullivan. Cross-Lingual Ontology Mapping — An Investigation of the Impact of Machine Translation. In *Proceedings of the 4th Asian Conference on The Semantic Web*, ASWC '09, pages 1–15. Springer-Verlag, 2009.

[37] Bo Fu, Rob Brennan, and Declan O'Sullivan. A Configurable Translation-Based Cross-Lingual Ontology Mapping System to adjust Mapping Outcome. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 15(3), 2012.

[38] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. URL `http://dl.acm.org/citation.cfm?id=1625275.1625535`.

[39] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.*, 34(1):443–498, March 2009. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=1622716.1622728`.

[40] Dimitrios Galanis and Ion Androutsopoulos. Generating Multilingual Descriptions from Linguistically Annotated OWL Ontologies: the NaturalOWL System. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 143–146, Saarbrücken, Germany, June 2007. DFKI GmbH. URL `http://www.aclweb.org/anthology/W07-2322`. Document D-07-01.

[41] Jianfeng Gao, Jian-Yun Nie, and Ming Zhou. Statistical query translation models for cross-language information retrieval. *ACM Transactions on*

Asian Language Information Processing , 5(4):323–359, December 2006. ISSN 1530-0226. URL http://doi.acm.org/10.1145/1236181.1236184.

[42] Fredric C. Gey, Noriko Kando, and Carol Peters. Cross-language information retrieval: the way ahead. *Information Processing & Management*, 41 (3):415 – 431, 2005. ISSN 0306-4573. URL http://dx.doi.org/10.1016/j.ipm.2004.06.006.

[43] Julio Gonzalo, Felisa Verdejo, Carol Peters, and Nicoletta Calzolari. Applying EuroWordNet to Cross-Language Text Retrieval. *Computers and the Humanities*, 32(2/3):185–207, 1998. URL http://www.jstor.org/stable/30200460.

[44] Jorge Gracia and Kartik Asooja. Monolingual and cross-lingual ontology matching with CIDER-CL: evaluation report for OAEI 2013. In *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21*, pages 109–116, 2013. URL http://ceur-ws.org/Vol-1111/oaei13_paper2.pdf.

[45] Jorge Gracia, Jordi Bernad, and Eduardo Mena. Ontology matching with CIDER: evaluation report for OAEI 2011. In Shvaiko et al. [116]. URL http://ceur-ws.org/Vol-814/oaei11_paper3.pdf.

[46] Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. Challenges for the Multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71, March 2012. ISSN 1570-8268. URL http://dx.doi.org/10.1016/j.websem.2011.09.001.

[47] Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gômez-Pérez. Cross-lingual Linking on the Multilingual Web of Data. In *Proc. of the 3rd Workshop on the Multilingual Semantic Web (MSW 2012) at ISWC 2012, Boston (USA)*. CEUR-WS.org, 2012.

[48] Gregory Grefenstette. *Cross-Lingual Information retrieval*. Kluwer Academic Publishers, 1998.

[49] Gunnar Aastrand Grimnes, Peter Edwards, and Alun D. Preece. Instance based clustering of semantic web resources. In *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, pages 303–317, 2008. URL http://dx.doi.org/10.1007/978-3-540-68234-9_24.

[50] Harry Halpin and Patrick J. Hayes. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010. URL `http://ceur-ws.org/Vol-628/ldow2010_paper09.pdf`.

[51] Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1192–1201, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-63-3. URL `http://dl.acm.org/citation.cfm?id=1699648.1699665`.

[52] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.

[53] Mauricio A. Hernández and Salvatore J. Stolfo. The Merge/Purge Problem for Large Databases. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, SIGMOD '95, pages 127–138, New York, NY, USA, 1995. ACM. ISBN 0-89791-731-6. URL `http://doi.acm.org/10.1145/223784.223807`.

[54] Sven Hertling and Heiko Paulheim. WikiMatch - Using Wikipedia for Ontology Matching. In *Proceedings of the 7th International Workshop on Ontology Matching*, volume 946 of *CEUR Workshop Proceedings*, Boston, MA, USA, November 2012. CEUR-WS.org.

[55] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, August 2009.

[56] Gail Hodge. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Technical report, Council on Library and Information Resources, Washington, DC. Digital Library Federation, 2000.

[57] Soon Gill Hong, Saemi Jang, Young Ho Chung, Mun Yong Yi, and Key-Sun Choi. Interlinking korean resources on the web. In *Semantic Technology, Second Joint International Conference, JIST 2012, Nara, Japan, December 2-4, 2012. Proceedings*, pages 382–387, 2012. URL `http://dx.doi.org/10.1007/978-3-642-37996-3_33`.

[58] Afraz Jaffri, Hugh Glaser, and Ian Millard. URI disambiguation in the context of linked data. In *Proceedings of the WWW2008 Workshop on Linked Data on the Web, LDOW 2008, Beijing, China*, 2008. URL `http://ceur-ws.org/Vol-369/paper19.pdf`.

[59] Saemi Jang, Satria Hutomo, Soon Gill Hong, and Mun Yong Yi. Interlinking Multilingual LOD Resources: A Study on Connecting Chinese, Japanese, and Korean Resources Using the Unihan Database. In *Proceedings of the ISWC 2013 Posters & Demonstrations Track, Sydney, Australia, October 23, 2013*, pages 229–232, 2013. URL `http://ceur-ws.org/Vol-1035/iswc2013_poster_13.pdf`.

[60] Heng Ji, Ralph Grishman, and Hoa Trang Dang. An Overview of the TAC2011 Knowledge Base Population Track. In *Proceesings of Text Analysis Conference (TAC)*, 2011.

[61] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000. ISBN 0130950696.

[62] Vipul Kashyap and Amit Sheth. *Information Brokering Across Heterogeneous Digital Data: A Metadata-based Approach*. Kluwer Academic Publishers, 2000.

[63] Kazuaki Kishida. Technical issues of cross-language information retrieval: A review. *Inf. Process. Manage.*, 41(3):433–455, 2005. ISSN 0306-4573. URL `http://dx.doi.org/10.1016/j.ipm.2004.06.007`.

[64] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the 10th Machine Translation Summit*, pages 79–86. AAMT, 2005. URL `http://mt-archive.info/MTS-2005-Koehn.pdf`.

[65] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. Technical report, World Wide Web Consortium, 1999.

[66] Tatiana Lesnikova. Interlinking RDF Data in Different Languages. The TOTh Workshop (Terminology and Ontology : Theories and applications), 2014. URL `http://porphyre.org/workshop-toth/2014-en`.

[67] Tatiana Lesnikova, Jérôme David, and Jérôme Euzenat. Interlinking English and Chinese RDF Data Sets Using Machine Translation. In Johanna

Völker, Heiko Paulheim, Jens Lehmann, Harald Sack, and Vojtech Svátek, editors, *Proceedings of the 3rd Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD 2014)*, volume 1243. CEUR-WS.org, 2014.

[68] Tatiana Lesnikova, Jérôme David, and Jérôme Euzenat. Interlinking English and Chinese RDF Data Using BabelNet. In Christine Vanoirbeek and Pierre Genevès, editors, *Proceedings of the 2015 ACM Symposium on Document Engineering, DocEng 2015, Lausanne, Switzerland, September 8-11, 2015*, pages 39–42. ACM, 2015. ISBN 978-1-4503-3307-8. URL `http://doi.acm.org/10.1145/2682571.2797089`.

[69] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*, 41(3):523 – 547, 2005. ISSN 0306-4573. URL `http://dx.doi.org/10.1016/j.ipm.2004.06.012`.

[70] Mingyang Li, Yao Shi, Zhigang Wang, and Yongbin Liu. Building a large-scale cross-lingual knowledge base from heterogeneous online wikis. In *Natural Language Processing and Chinese Computing - 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings*, pages 413–420, 2015. URL `http://dx.doi.org/10.1007/978-3-319-25207-0_37`.

[71] Feiyu Lin and Andrew Krizhanovsky. Multilingual Ontology Matching Based on Wiktionary Data Accessible via SPARQL Endpoint. In *Proceedings of the 13th Russian Conference on Digital Libraries*, RCDL'2011, pages 19–26, Voronezh, Russia, October 19-22 2011.

[72] Michael Littman, Susan T. Dumais, and Thomas K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-Language Information Retrieval, chapter 5*, pages 51–62. Kluwer Academic Publishers, 1998.

[73] Prasenjit Majumder, Mandar Mitra, Pushpak Bhattacharyya, L. Venkata Subramaniam, Danish Contractor, and Paolo Rosso, editors. *Multilingual Information Access in South Asian Languages, Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*, volume 7536 of *Lecture Notes in Computer Science*, 2013. Springer. ISBN 978-3-642-40086-5. URL `http://dx.doi.org/10.1007/978-3-642-40087-2`.

[74] Ranbeer Makin, Nikita Pandey, Prasad Pingali, and Vasudeva Varma. Approximate string matching techniques for effective clir among indian languages. In Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, editors, *WILF*, volume 4578 of *Lecture Notes in Computer Science*, pages 430–437. Springer, 2007. ISBN 978-3-540-73399-7. URL http://dblp.uni-trier.de/db/conf/wilf/wilf2007.html#MakinPPV07.

[75] J. Scott McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 208–214, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. URL http://dx.doi.org/10.3115/1034678.1034716.

[76] John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. Interchanging lexical resources on the semantic web. In *Language Resources and Evaluation*, volume 46(4), pages 701–719. Springer, 2012.

[77] Paul McNamee and James Mayfield. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2): 73–97, 2004. ISSN 1386-4564. URL http://dx.doi.org/10.1023/B:INRT.0000009441.78971.be.

[78] Christian Meilicke, Cássia Trojahn dos Santos, Ondřej Šváb-Zamazal, and Dominique Ritze. Multilingual Ontology Matching Evaluation - A First Report on Using MultiFarm. In Elena Simperl, Barry Norton, Dunja Mladenic, Emanuele Della Valle, Irini Fundulaki, Alexandre Passant, and Raphaël Troncy, editors, *The Semantic Web: ESWC 2012 Satellite Events - ESWC 2012 Satellite Events, Heraklion, Crete, Greece, May 27-31, 2012. Revised Selected Papers*, volume 7540 of *Lecture Notes in Computer Science*, pages 132–147. Springer, 2012. ISBN 978-3-662-46640-7. URL http://dx.doi.org/10.1007/978-3-662-46641-4_10.

[79] Christian Meilicke, Raúl García-Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondřej Šváb Zamazal, Vojtěch Svátek, Andrei Tamilin, Cássia Trojahn, and Shenghui Wang. MultiFarm: A Benchmark for Multilingual Ontology Matching. *Journal of Web Semantics*, 15:62–68, 2012.

[80] Chris Mellish and Xiantang Sun. The semantic web as a linguistic resource: Opportunities for natural language generation. *Knowl.-Based Syst.*, 19(5):

298–303, 2006. URL http://dx.doi.org/10.1016/j.knosys.2005.11.011.

[81] Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/, 2009.

[82] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

[83] Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)*. Oxford University Press, 2003. ISBN 0198238827.

[84] Sean Monahan, John Lehmann, Timothy Nyberg, Jesse Plymale, and Arnold Jung. Cross-Lingual Cross-Document Coreference with Entity Linking. In *Proceedings of the Text Analysis Conference (TAC)*. NIST, 2011. URL http://www.nist.gov/tac/publications/2011/papers.html.

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

[86] Ahsan Morshed, Caterina Caracciolo, Gudrun Johannsen, and Johannes Keizer. Thesaurus Alignment for Linked Data publishing. In *International Conference on Dublin Core and Metadata Applications*, pages 37–46, 2011.

[87] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1): 32–38, 1957.

[88] Vivi Nastase and Michael Strube. Transforming Wikipedia into a Large Scale Multilingual Concept Network. *Artif. Intell.*, 194:62–85, January 2013. ISSN 0004-3702. URL http://dx.doi.org/10.1016/j.artint.2012.06.008.

[89] Vivi Nastase, Michael Strube, Benjamin Börschinger, Cäcilia Zirn, and Anas Elghafari. Wikinet: A very large scale multi-lingual concept network. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1015–1022, 2010.

[90] Roberto Navigli and Simone Paolo Ponzetto. Babelrelate! A joint multilingual approach to computing semantic relatedness. In *Proceedings of*

the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada, 2012. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5112.

[91] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250, 2012.

[92] Axel-Cyrille Ngonga Ngomo and Sören Auer. LIMES: A time-efficient approach for large-scale link discovery on the web of data. In *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2312–2317. AAAI Press, 2011.

[93] Khai Nguyen, Ryutaro Ichise, and Bac Le. SLINT: a schema-independent linked data interlinking system. In Pavel Shvaiko, Jérôme Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Fridman Noy, and Heiner Stuckenschmidt, editors, *Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012*, volume 946 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012. URL http://ceur-ws.org/Vol-946/om2012_Tpaper1.pdf.

[94] Xing Niu, Shu Rong, Yunlong Zhang, and Haofen Wang. Zhishi.links results for OAEI 2011. In Shvaiko et al. [116]. URL http://ceur-ws.org/Vol-814/oaei11_paper16.pdf.

[95] Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, AMTA '98, pages 472–483, London, UK, UK, 1998. Springer-Verlag. ISBN 3-540-65259-0. URL http://dl.acm.org/citation.cfm?id=648179.749060.

[96] Douglas W. Oard. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):79–84, 2003. ISSN 1530-0226. URL http://doi.acm.org/10.1145/974740.974741.

[97] Aris M. Ouksel and Amit Sheth. Semantic Interoperability in Global Information Systems. *SIGMOD Rec.*, 28(1):5–12, March 1999. ISSN 0163-5808. URL http://doi.acm.org/10.1145/309844.309849.

[98] Andreas Paepcke, Chen-Chuan K. Chang, Terry Winograd, and Héctor García-Molina. Interoperability for Digital Libraries Worldwide. *Commun. ACM*, 41(4):33–42, April 1998. ISSN 0001-0782. URL http://doi.acm.org/10.1145/273035.273044.

[99] Carol Peters and Paraic Sheridan. Multilingual information access. In *Proceedings of the Third European Summer-School on Lectures on Information Retrieval-Revised Lectures*, ESSIR'00, pages 51–80, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-41933-0. URL `http://dl.acm.org/citation.cfm?id=646171.678755`.

[100] Ari Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'98, pages 55–63, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. URL `http://doi.acm.org/10.1145/290941.290957`.

[101] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[102] Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Cross-language plagiarism detection. *Lang. Resour. Eval.*, 45(1):45–62, March 2011. ISSN 1574-020X. URL `http://dx.doi.org/10.1007/s10579-009-9114-z`.

[103] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Käsper, and Irina Temnikova. Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING'04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1220355.1220493`.

[104] Shelley Powers. *Practical RDF*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2003. ISBN 0596002637.

[105] Yuzhong Qu, Wei Hu, and Gong Cheng. Constructing virtual documents for ontology matching. In *Proceedings of the 15th International Conference on World Wide Web*, pages 23–31, Edinburgh, Scotland, May 23 - 26 2006. ACM Press, New York, NY.

[106] Razieh Rahimi, Azadeh Shakery, and Irwin King. Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework. *Information Processing & Management*, 2015. ISSN 0306-4573. URL `http://dx.doi.org/10.1016/j.ipm.2015.08.001`.

[107] Reinhard Rapp, Serge Sharoff, and Bogdan Babych. Identifying word translations from comparable documents without a seed lexicon. In *Pro-*

ceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, pages 460–466, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/888.html.

[108] Gerard Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[109] François Scharffe and Jérôme Euzenat. Linked Data Meets Ontology Matching - Enhancing Data Linking through Ontology Alignments. In *KEOD 2011 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Paris, France*, pages 279–284, 2011.

[110] François Scharffe, Yanbin Liu, and Chuguang Zhou. RDF-AI: an architecture for RDF datasets matching, fusion and interlink. In *Proc. IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena (CA US)*, 2009.

[111] Gilles Sérasset. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361, 2015. URL http://dx.doi.org/10.3233/SW-140147.

[112] Gilles Sérasset and Andon Tchechmedjiev. Dbnary: Wiktionary as linked data for 12 language editions with enhanced translation relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, LREC 2014*, pages 68–71, 2014.

[113] Chao Shao, Linmei Hu, and Juanzi Li. RiMOM-IM results for OAEI 2014. In *Proceedings 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*, pages 149–154, 2014. URL http://ceur-ws.org/Vol-1317/oaei14_paper7.pdf.

[114] Mohamed Ahmed Sherif and Axel-Cyrille Ngonga Ngomo. Semantic Quran: A Multilingual Resource for Natural-Language Processing. *Semantic Web Journal*, 6(4):339–345, 2015. URL http://www.semantic-web-journal.net/system/files/swj503_0.pdf.

[115] Amit P. Sheth. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In M.F. Goodchild, M.J. Egenhofer, R. Fegeas, and C.A. Kottman, editors, *Interoperating Geographic Information Systems*, volume 495, pages 5–29. Kluwer, 1999.

[116] Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz, editors. *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, volume 814 of *CEUR Workshop Proceedings*, 2011. CEUR-WS.org. URL `http://ceur-ws.org/Vol-814`.

[117] Agnès Simon, Romain Wenz, Vincent Michel, and Adrien Di Mascio. Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF (French National Library). In *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 563–577. Springer, 2013.

[118] Anestis Sitas and Sarantos Kapidakis. Duplicate detection algorithms of bibliographic descriptions. *Library Hi Tech*, 26(2):287–301, 2008. URL `http://dx.doi.org/10.1108/07378830810880379`.

[119] Philipp Sorg and Philipp Cimiano. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*, 2008.

[120] Philipp Sorg and Philipp Cimiano. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems*, NLDB'09, pages 36–48, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-12549-2, 978-3-642-12549-2. URL `http://dx.doi.org/10.1007/978-3-642-12550-8_4`.

[121] Philipp Sorg and Philipp Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26 – 45, 2012. ISSN 0169-023X. URL `http://dx.doi.org/10.1016/j.datak.2012.02.003`.

[122] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984. ISBN 0-201-14472-7.

[123] Dennis Spohr, Laura Hollink, and Philipp Cimiano. A Machine Learning Approach to Multilingual and Cross-lingual Ontology Matching. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, pages 665–680. Springer-Verlag, 2011.

[124] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A multilingual

aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, 2006.

[125] Mark Stevenson and Paul D. Clough. Eurowordnet as a resource for cross-language information retrieval. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, 2004. URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/76.pdf`.

[126] Stephanie Strassel, Mike Maxwell, and Christopher Cieri. Linguistic resource creation for research and technology development: A recent experiment. *ACM Transactions on Asian Language Information Processing*, 2 (2):101–117, 2003. ISSN 1530-0226. URL `http://doi.acm.org/10.1145/974740.974743`.

[127] Xiantang Sun and Chris Mellish. An Experiment on "Free Generation" from Single RDF Triples. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG'07, pages 105–108, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1610163.1610181`.

[128] Cássia Trojahn, Paulo Quaresma, and Renata Vieira. An API for Multilingual Ontology Matching. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3830–3835, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[129] Mark van Assem, Aldo Gangemi, and Guus Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the LREC (2006)*, pages 237–242, 2006.

[130] Daniel Vila-Suero, Boris Villazón-Terrazas, and Asunción Gómez-Pérez. datos.bne.es: a library linked dataset. *Semantic Web Journal*, 4(3):307–313, 2012.

[131] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and Maintaining Links on the Web of Data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC'09, pages 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.

[132] Piek Vossen and Computer Centrum Letteren. Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7, 1997.

[133] Shenghui Wang, Antoine Isaac, Balthasar A. C. Schopman, Stefan Schlobach, and Lourens van der Meij. Matching Multilingual Subject Vocabularies. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, volume 5714, pages 125–137. Springer, Heidelberg, 2009.

[134] Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 459–468, 2012. URL `http://doi.acm.org/10.1145/2187836.2187899`.

[135] Zhichun Wang, Zhigang Wang, Juanzi Li, and Jeff Z. Pan. Knowledge extraction from Chinese wiki encyclopedias. *Journal of Zhejiang University - Science C*, 13(4):268–280, 2012. URL `http://dx.doi.org/10.1631/jzus.C1101008`.

[136] Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. In *International Semantic Web Conference (Posters & Demos)*, volume 1035 of *CEUR Workshop Proceeding*, pages 121–124. CEUR-WS.org, 2013.

[137] Yunqing Xia, Guoyu Tang, Peng Jin, and Xia Yang. CLTC: A Chinese-English Cross-lingual Topic Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 532–537, 2012.

[138] Benjamin Zapilko, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences. *Semantic Web journal (SWJ)*, 4(3):257–263, 2013.

[139] Lihua Zhao and Ryutaro Ichise. Instance-Based Ontological Knowledge Acquisition. In *Proc.10th International Conference, ESWC 2013*, volume 7882, pages 155–169. Springer Berlin Heidelberg, 2013.

[140] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45(1):1:1–1:44, December 2012. ISSN 0360-0300. URL `http://doi.acm.org/10.1145/2379776.2379777`.