Cross-Lingual Natural Language Querying over the Web of Data

Nitish Aggarwal[1], Tamara Polajnar[2], and Paul Buitelaar[1]

[1] Unit for Natural Language Processing, Digital Enterprise Research Institute,
National University of Ireland, Galway
[2] Computer Laboratory, University of Cambridge, Cambridge
firstname.lastname@deri.org

Abstract. The rapid growth of the Semantic Web offers a wealth of semantic knowledge in the form of Linked Data and ontologies, which can be considered as large knowledge graphs of marked up Web data. However, much of this knowledge is only available in English, affecting effective information access in the multilingual Web. A particular challenge arises from the vocabulary gap resulting from the difference in the query and the data languages. In this paper, we present an approach to perform cross-lingual natural language queries on Linked Data. Our method includes three components: entity identification, linguistic analysis, and semantic relatedness. We use Cross-Lingual Explicit Semantic Analysis to overcome the language gap between the queries and data. The experimental results are evaluated against 50 German natural language queries. We show that an approach using a cross-lingual similarity and relatedness measure outperforms other systems that use automatic translation. We also discuss the queries that can be handled by our approach.

Keywords: Semantic Web, Natural Langauge Querying, CLIR.

1 Introduction

1.1 Motivation

In the last decade, the Semantic Web community has been working extensively towards creating standards, which tend to increase the accessibility of available information on the Web, by providing structured metadata[1]. Yahoo research recently reported [1] that 30% of all HTML pages on the Web contain structured metadata such as microdata, RDFa, or microformat. This structured metadata facilitates the possibility of automatic reasoning and inferencing. Thus, by embedding such knowledge within web documents, additional key information about the semantic relations among data objects mentioned in the web pages can be captured.

One of the most difficult challenge in multilingual web research is cross-lingual document retrieval, i.e. retrieval of relevant documents that are written in a language other than the query language. To address this issue we present a method for cross-lingual

[1] http://events.linkeddata.org/ldow2012/slides/Bizer-LDOW2012-Panel-Background-Statistics.pdf

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 152–163, 2013.

natural language querying, which aims to retrieve all relevant information even if it is only available in a language different from the query language. Our approach differs from the state-of-the-art methods, which mainly consist of translating the queries into document languages ([2], [3]). However, the poor accuracy of automatic translation of short texts like queries makes this approach problematic. Hence, using large knowledge bases as an interlingua [4] may prove beneficial. The approach discussed here considers Linked Data as a structured knowledge graph. The Linked Open Data (LOD) cloud currently contains more than 291 different structured knowledge repositories in RDF[2] format, which are linked together using “DBpedia”, “freebase” or “YAGO”. It contains a large number of instances in many different languages, however, the vocabulary used to define ontology relations is mainly in English. Thus, querying this knowledge base is not possible in other languages even if the instances are multilingual. Cross-lingual natural language querying is required to access this structured knowledge base, which is the main objective of our approach.

1.2 Problem

Retrieval of structured data, in general, requires structured queries; however, effective construction of such queries is a laborious process. In order to provide a flexible querying environment, we propose to automatically construct a structured query from a natural language query (NL-Query). While there are several efforts ([5], [6], [7]) to convert a NL-Queries into structured SPARQL[3] queries in the monolingual scenario, the multilingual scenario offers further challenges. For example, the problem of mapping the query vocabulary to the ontology vocabulary is exacerbated by poor quality of automatic translation for short text and by the lack of multilingual structured resources. Therefore, to avoid relying on automatic translation, we present a novel approach for cross-lingual NL-Query formulation, which includes entity search, linguistic analysis, and semantic similarity and relatedness measure. We used Cross-Lingual Explicit Semantic Analysis (CL-ESA) to calculate the semantic relatedness scores between vocabularies in different languages.

1.3 Contribution

The main focus of our approach is the interpretation of NL-Queries by traversal over the structured knowledge graph, and the construction of a corresponding SPARQL query. As discussed in Section 1.2, translation based approaches for cross-lingual NL-Queries suffer from the poor quality of automatic translation. Therefore, in this paper, we introduce a novel approach for performing cross-lingual NL-Queries over structured knowledge base, without automatic translation. As an additional contribution, we have created and analyzed a benchmark dataset of 50 NL-Queries in German. We discuss the results of a comparison of our method with an automatic translation method over the 28 NL-Queries that can be addressed by our approach.

[2] Resource Description Framework (RDF) is the World Wide Web consortium (W3C) specification to represents the conceptual description. It was designed as a metadata data model.

[3] `http://www.w3.org/TR/rdf-sparql-query/`

Our algorithm can also be used for cross-lingual document retrieval provided that the document collection is already marked up with a knowledge base, for instance, Wikipedia articles annotated with DBpedia.

2 State of the Art

Most of the proposed approaches that address the task of Cross-Lingual Information Retrieval (CLIR) reduce the problem into a monolingual scenario by translating the search query or documents in the corresponding language. Many of them perform query translation ([8], [9], [2], [3])) into the language of the documents. However, all of these approaches suffer from the poor performance of machine translation on short texts (query). Jones et al. [3] performed query translation by restricting the translation to the cultural heritage domain, while Nquyen et al. [2] makes use of the Wikipedia cross-lingual links structure.

Without relying on machine translation, some approaches ([10], [11], [12]) make use of distributional semantics. They calculate a cross-lingual semantic relatedness score between the query and the documents. However, none of these approaches take any linguistic information into account, and do not make use of large available structured knowledge bases. With the assumption that documents of different languages are already marked-up with the knowledge base (for instance, Wikipedia articles are annotated with DBpedia), the problem of CLIR can be converted into querying over structured data. There is still a language barrier, as queries can be in different languages, while most of the structured data is only available in English. Qall-Me [13] performs NL-Querying over structured information by using textual entailment to convert a natural language question into SPARQL. This system relies on availability of multilingual structured data. It can only retrieve the information that is available in the query language. Therefore, this system is not able to perform CLIR. Freitas et al. [5] proposed an approach for natural language querying over linked data, based on the combination of entity search, a Wikipedia-based semantic relatedness (using ESA) measure, and spreading activation. Their approach is similar to ours, but it can not deal with different languages.

3 Background

3.1 DBpedia and SPARQL

We used DBpedia[4] as a knowledge base for our experiments. DBpedia is a large structured knowledge base, which is extracted from Wikipedia info-boxes. It contains data in the form of a large RDF graph, where each node represents an entity or a literal and the edges represent relations between entities. Each RDF statement can be divided into a subject, predicate and object. DBpedia contains a large ontology, describing more than 3.5 millions instances, forming a large general structured knowledge source. Also, it is very well-connected to several other Linked Data repositories in the Semantic Web. As

[4] http://dbpedia.org/

DBpedia instances are extracted from Wikipedia, they contains cross-links across the different languages, however, the properties (or relations) associated with the instances, are mainly defined in English.

In order to query DBpedia, a structured query is required. SPARQL is the standard structured query language for RDF, and allows users to write unambiguous queries to retrieve RDF triples.

3.2 Cross-Lingual ESA

Semantic relatedness of two given terms can be obtained by calculating the similarity between two high dimensional vectors in a distributed semantic space model (DSM). According to the distributional hypothesis, the semantic meaning of a word can (at least to a certain extent) be inferred from its usage in context, that is its distribution in text. This semantic representation is built through a statistical analysis over the large contextual information in which a term occurs. One recent popular model to calculate this relatedness by using the distributed semantics is Explicit Semantic Analysis (ESA) proposed by Gabrilovich and Markovitch [14], which attempts to represent the semantics of a given term in a high dimensional vector of explicitly defined concepts. In the original paper the Wikipedia articles were used to built the ESA model. Every dimension of the high dimensional vector reflects a unique Wikipedia concept or title, and the weight of the dimensions are created by taking the TF-IDF weight of a given term in the corresponding title of a Wikipedia document.

An interesting characteristic of Wikipedia is that this very large collective knowledge is available in multiple languages, which facilitates an extension of existing ESA for multiple languages called Cross-Lingual Explicit Semantic Analysis (CL-ESA) proposed by Sorg et al. [15]. The articles in Wikipedia are linked together across the languages. This cross-lingual linked structure can provide a mapping of a vector in one language to another. To understand CLESA, let us take two terms t_s in language L_s and t_t in language L_t. As a first step, a concept vector for t_s is created using the Wikipedia corpus in L_s. Similarly, the concept vector for t_t is created in L_t. Then, one of the concept vectors can be converted to the other language by using the cross-lingual links between articles across the languages, provided by Wikipedia. After obtaining both of the concept vectors in one language, the relatedness of the terms t_s and t_t can be calculated by using the cosine product, similar to ESA. For better efficiency, we chose to make a multilingual index by composing poly-lingual Wikipedia articles using the cross-lingual mappings. In such a case, no conversion of the concept vector in one language to the other is required. Instead, it is possible to represent the Wikipedia concept with some unique name common to all languages such as, for instance, the Uniform Resource Identifier (URI) of the English Wikipedia.

4 Approach

The key to our approach is the interpretation of NL-Queries in different languages, by using a combination of *entity identification, linguistic analysis* and *cross-lingual similarity and relatedness measure*. Figure 1 shows the three components of our approach along with an example of a NL-Query in German[5]. The interpretation process starts with the identification of possible entities appearing in a given NL-Query, followed by linguistic analysis of the NL-Query. The system performs the whole pipeline with all of the identified entities and takes union over all of the retrieved results. Using the dependencies provided by the linguistic analysis, our system determines the next term that will be compared with all the relations associated with the identified entity, to find the best matched relation. For instance, in example shown in Figure 1, the system identified "Bill Clinton" as entity and "Tochter" as next term. Following the process, it calculates the similarity score with every relation associated with "Bill Clinton" and finds the maximum similar relation to obtain the next entity from the knowledge base.

4.1 Entity Identification

The first step of the interpretation process is the identification of potential entities, i.e. the Linked Data concepts (classes and instances), present in the NL-Query. A baseline entity identification can be defined as the identification of an exact match between the label of a concept against the term appearing in the NL-Query; for example, DBpedia: Bill_Clinton shown in Figure 1. "Bill Clinton" is the name of a person and it appeared as a label of DBpedia: Bill_Clinton URI in the database. However, a term such as "Ministerpräsidenten von Spanien" and "Christus im Sturm auf dem See Genezareth" do not appear as labels in the database. Therefore, in order to resolve this issue, we translate the term to get the approximate term in the corresponding language and find the best matched label in the database. We use the Bing translation system[6] to perform the automatic translation but the quality of translation is not very promising and we do not get the exact translation of a given label. Therefore, we calculate the token edit distance between translated label and the labels in our database and select the maximum matched one. For instance, the translation of "Christ in the storm on the sea of Galilee" is "Christ in storm on the sea of Galilee" but label of the appropriate concept is "The Storm on the Sea of Galilee".

In addition, our approach includes a disambiguation process, in which we disambiguate the selected concept candidates based on their associated relations in the knowledge base. For instance, in a given NL-Query "Wie viele Angestellte hat Google?"@de[7] two different DBpedia entities can be found with the label "Google", i.e. "DBpedia: Google_Search" and "DBpedia: Google". We calculate similarity scores with all associated relations of both, and find that term "Angestellte" in the NL-Query obtained maximum similarity score with the relation "number Employees", which is associated with "DBpedia: Google".

[5] Translated from the QALD-2 challenge dataset, which has 100 NL-Queries in English, over DBpedia.

[6] `http://www.bing.com/translator`

[7] Translation of "How many employees does Google have?" from the English test dataset.

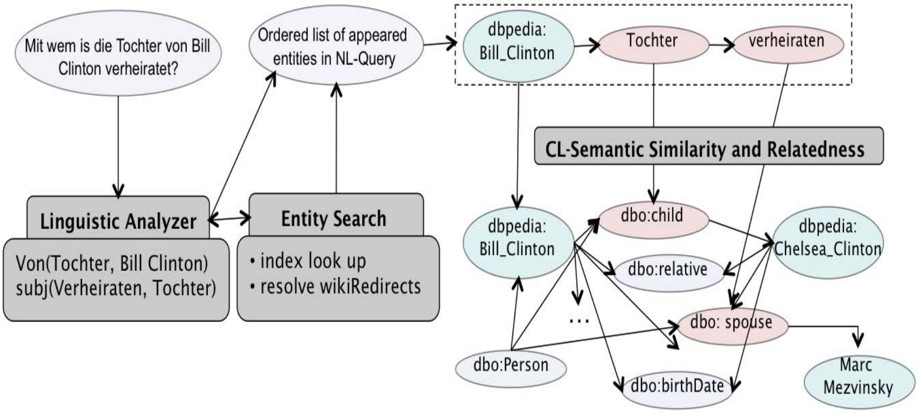

Fig. 1. Query interpretation pipeline for the German NL-Query "Mit wem is die Tochter von Bill Clinton verheiratet?" ("Who is the daughter of Bill Clinton married to?"@en)

4.2 Linguistic Analysis

Linguistic analysis of the NL-Queries is needed to get the dependencies among the identified entities and terms. We use the Stanford parser[8] for German to generate the dependencies. Following these dependencies, we convert the given NL-Query into a Direct Acyclic Graph (DAG). Vertices of the generated DAG represent the entities and edges reflect the terms directly dependent on the obtained entities (vertices). Figure 1 shows the DAG obtained from our example query "Mit wem is die Tochter von Bil Clinton verheiratet?". To generate the DAG, first we obtain the central entity from the previous step. With the relations of this central entity, semantic matching will be performed. Therefore, we retrieve the directly dependent terms of the central entity by following the generated Stanford typed dependencies, and add them into the DAG. Similarly, we perform this action for all the other terms in the list. For instance, in our example NL-Query shown in Figure 1, firstly, the system identifies "Bill Clinton" as a central entity,[9] and then "Tochter" as direct dependent of "Bill Clinton" followed by "verheiratet" as direct dependent of "Tochter".

4.3 Knowledge Graph Traversing Using Semantic Similarity and Relatedness

A knowledge graph can be defined as the structured data of well-connected entities and their relations. Our next step is to find such relations of selected central entity in the knowledge base that are best matches with the term directly dependent on this central entity in the generated DAG. First, we search for the entity "Bill Clinton" in DBpedia as our approach takes DBpedia as knowledge base, and retrieve all of the relations (DBpedia properties) associated with it. Then, we find the best semantically match DBpedia

[8] `http://nlp.stanford.edu/software/lex-parser.shtml`

[9] The term to start the search around in the whole DBpedia graph.

property of the direct dependent term "Tochter" by calculating a cross-lingual similarity score between all the DBpedia properties of Bill Clinton and "Tochter". After obtaining the relevant property, i.e. "child", we find the entity DBpedia: Chelsea_Clinton, connected with entity "DBpedia: Bill_Clinton" by property "child". We perform the same steps with the retrieved entity for the directly dependent term "verheiratet" of "Tochter", and so on until the end of the DAG. Finally, we retrieve the most relevant entity and all the associated documents in different languages containing a description about this entity.

5 Evaluation

5.1 Datasets

In order to evaluate our approach, we created a testset of 50 NL-Queries in German. The benchmark is created by manually translating the English NL-Queries provided by the "Question Answering over Linked Data (QALD-2)" dataset, consists of 100 NL-Queries in English over DBpedia. All of the NL-Queries are annotated with keywords, corresponding SPARQL queries and answers retrieved from DBpedia. Also, every NL-Query specifies some additional attributes, for example, if a mathematical operation such as aggregation, count or sort is needed in order to retrieve the appropriate answers.

Table 1. Query categorization of training and test dataset

Dataset	Simple	Template-based	SPARQL aggregation
Training	27	11	12
Test	28	10	12

We translated QALD-2 dataset and divided it into two parts, one for training and one for testing. Therefore, each dataset contains 50 NL-Queries in German. We performed a manual analysis to keep the same complexity level in both the datasets. We divided all of the NL-Queries into three different categories: simple, template-based and SPARQL aggregation. Simple queries contain the DBpedia entities and their relations (DBpedia properties), and do not need a predefined template or rule to construct the corresponding SPARQL query. However, these queries include semantic and linguistic variations, that means they express the DBpedia properties by using related terms rather than having the exact label of a property. For instance, in a given query "How tall is Michael Jordan?", "tall" does not appear in the vocabulary of DBpedia properties, however, the answer of the query can be retrieved by DBpedia property "height" appearing with "DBpedia: Michael_Jordan". Those queries, which required predefined templates or rules, are categorized as template-based [6] queries, for example, the query "Give me all professional skateboarders from Sweden." required a predefined template for retrieving all persons with occupation Skateboarding and born in Sweden. SPARQL aggregation type of queries need performing a mathematical operation such as aggregation, count or sort, therefore, they also require a predefined template.

Following the categorization, we divided the dataset into two parts by keeping an equal number of queries in each category. We then performed our experiments on the prepared test dataset of 50 NL-Queries in German. Table 1 shows the statistics about both the datasets. We are extending these datasets for other languages and they are freely available.

Table 2. Error type and its distribution over 50 natural language queries and 28 selected natural language queries in German

Error Type	**No of NL-Queries**	
	out of 50	**out of 28**
Entity Identification without Translation	10	3
Entity Identification with Translation	7	1
Linguistic Analysis	14	4
At least one	18	5

5.2 Experiment

We evaluated the outcome of our approach at all three stages of the processing pipeline: 1) entity identification, 2) linguistic analysis, and 3) semantic similarity and relatedness measures. This way, we can investigate the errors introduced by individual components. As shown in Figure 1, the third component "semantic similarity and relatedness measures" relies on the correctness of the constructed DAG, i.e. on the performance of both the previous components (entity identification and linguistic analysis). Therefore, it is important to examine the performance of individual components. We evaluated the outcome of entity identification and linguistic analysis on all 50 NL-Queries of the test dataset. However, all of the template-based and SPARQL aggregation type NL-Queries are out of the scope of our settings. Therefore, we discuss the results obtained for remaining 28 NL-Queries. The entity identification component was evaluated in both ways; entity identification without using automatic translation and entity identification with automatic translation[10]. Table 2 shows that appropriate entities could not be found in 10 NL-Queries out of 50 NL-Queries and 3 NL-Queries out of 28. However, by using automatic translation the error is reduced to 7 and 1 NL-Queries respectively. To evaluate the performance of the linguistic analysis component, we counted the number of NL-Queries, for which the Stanford parser was unable to generate the dependencies. The statistics of the errors in linguistic analysis are shown in Table 2. As explained in Section 4.3, to find the relevant properties associated with the selected DBpedia entity, a comparison of all the properties and the next term from the DAG is needed. This requires a good cross-lingual similarity and relatedness measure. Therefore, to examine the effect of similarity and relatedness measure over automatic translation, we used three different settings in calculating the scores: a) automatic translation followed by

[10] The automatic translation was only used for those entities, that could not be found in the database with the given labels.

Table 3. Evaluation on 28 German NL-Queries

NL-Queries in German and English	**Translation with edit dist.**			**Tranlation with ESA**			**CL-ESA**		
	P	R	F1	P	R	F1	P	R	F1
1. Wer war der Nachfolger von John F. Kennedy?@de Who was the successor of John F. Kennedy?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2. Wie viele Studenten hat die Freie Universität Amsterdam?@de How many students does the Free University in Amsterdam have?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3. Über welche Länder erstreckt sich das Himalaya-Gebirgssystem?@de To which countries does the Himalayan mountain system extend?@en	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4. Gib mir alle Mitglieder von The Prodigy.@de Give me all members of Prodigy.@en	0.0	0.0	0.0	1.0	0.28	0.44	1.0	0.28	0.44
5. Wie groß ist Michael Jordan?@de How tall is Michael Jordan?@en	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
6. Wer ist der Gouverneur von Texas?@de Who is the governor of Texas?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7. Sean Parnell ist der Gouverneur welches US-Bundesstaates?@de Sean Parnell is the governor of which U.S. state?@en	0.33	1.0	0.5	0.33	1.0	0.5	0.33	1.0	0.5
8. Welches ist der Geburtsname von Angela Merkel?@de What is the birth name of Angela Merkel?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9. Wie oft hat Nicole Kidman geheiratet?@de How often did Nicole Kidman marry?@en	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
10. Wer hat Skype entwickelt?@de Who developed Skype?@en	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0
11. Gib mir alle Partnerstädte von Brünn.@de Give me all sister cities of Brno.@en	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
12. Wer hat Intel gegründet?@de Who founded Intel?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
13. Gib mir alle Rassen des Deutscher Schäferhund.@de Give me all breeds of the German Shepherd dog.@en	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
14. Wer hat den Reissverschluss erfunden?@de Who invented the zipper?@en	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
15. Welche Länder sind durch den Rhein verbunden?@de Which countries are connected by the Rhine?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
16. In welcher britischen Stadt ist der Hauptsitz des MI6?@de In which UK city are the headquarters of the MI6?@en	0.5	1.0	0.66	0.5	1.0	0.66	0.5	1.0	0.66
17. Welches sind die Spitznamen von San Francisco?@de What are the nicknames of San Francisco?@en	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
18. Gib mir die Astronauten von Apollo 14.@de Give me the Apollo 14 astronauts.@en	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19. Wie viele Kinder hatte Benjamin Franklin?@de Which ships were called after Benjamin Franklin?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
20. Welche Instrumente hat John Lennon gespielt?@de Which instruments did John Lennon play?@en	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
21. Wie viele Angestellte hat Google?@de How many employees does Google have?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
22. Wann ist Michael Jackson gestorben?@de When did Michael Jackson die?@en	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23. Wie hoch ist der Mount Everest?@de How high is the Mount Everest?@en	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24. Mit wem is die Tochter von Bill Clinton verheiratet?@de Who is the daughter of Bill Clinton married to?@en	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
24. Wer hat die Musik für Harold und Maude komponiert?@de Who composed the music for Harold and Maude?@en	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
26. Wo ist die Residenz des Ministerpräsidenten von Spanien?@de Where is the residence of the prime minister of Spain?@en	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
27. Aus welchem Land kommt der Schöpfer von Nijntje?@de Which country does the creator of Miffy come from?@en	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
28. Wer malte Christus im Sturm auf dem See Genezareth?@de Who painted The Storm on the Sea of Galilee?@en	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
Combined	**0.387**	**0.429**	**0.407**	**0.601**	**0.617**	**0.609**	**0.815**	**0.831**	**0.823**

string edit distance, b) automatic translation followed by monolingual Explicit Semantic Analysis (ESA), and c) Cross-Lingual Explicit Semantic Analysis (CL-ESA). To evaluate the performance of automatic translation over CL-ESA, we reduce the problem into a monolingual scenario, by translating the properties into corresponding language, in settings a and b. Automatic translation is not performed on the full text of a NL-Query but only on the properties because the quality of translation is not good enough to get the correct linguistic dependencies by using Stanford parser.

In first setting, we perform the translation and check if we can find the translated term in the listed properties by using Levenshtein edit distance approximation. While, in the second one, we calculate similarity and relatedness scores using ESA after performing automatic translation, to investigate if the automatic translation and semantic relatedness can complement each other. We do not use automatic translation in the third setting "cross-lingual semantic similarity and relatedness measure", but only rely on the scores generated by CL-ESA. The quality of final results generated by all three settings are analyzed manually and shown in Table 3 and we discuss it in detail in next Section 5.3.

5.3 Results and Discussion

Table 3 compares the results obtained by using three different settings of our approach: a) automatic translation followed by Levenshtein edit distance, b) automatic translation followed by monolingual ESA, and c) CL-ESA. It shows that automatic translation can not bridge the vocabulary gap between NL-Queries and DBpedia. That means there are large lexical variations in defining the relations of entities. Further, we investigate if automatic translation and monolingual ESA can complement each other. Although, the overall score generated by the combination of both is improved significantly over the score obtained by using automatic translation, the best results are generated by CL-ESA. The reason may be that combined errors introduced by using both translation and ESA is more than the error generated by CL-ESA.

Table 2 shows that 5 NL-Queries out of these 28 NL-Queries pose at least one type of error (entity identification or linguistic analysis), meaning that DAGs can be generated only for 23 NL-Queries out of 28. However, to reduce this error, we consider that keywords appearing in a given NL-Query may depend on the selected entity. For instance, the Stanford parser failed to generate the correct dependencies for Q28 and Q5 (listed in Table 3) but by considering the terms "groß" and "malte" to be dependent on the identified entities "Michael Jordan" and "Christus im Sturm auf dem See Genezareth" respectively, we could generate the correct DAGs. Therefore, we can test the third component of our approach on 25 NL-Queries out of 28 as we got the correct DAGs of 25 NL-Queries. We can see in Table 3, in our approach, by using translation we can retrieve the correct answers for 10 NL-Queries and partially correct for 2 NL-Queries; by using translation with the combination of ESA we can retrieve the correct answers for 15 NL-Queries and partially correct for 3 NL-Queries; by using CL-ESA we can retrieve the correct answers for 21 NL-Queries and partially correct for 3 NL-Queries.

In case of Q10, automatic translation followed by ESA failed because when the system tried to find the maximum related property with the term "developed" (translation of "entwickelt"), it obtained a higher ESA score for another property "operating system" than "developer". Our approach simply failed to find the results for Q14, due to the appearance of more than one highly related properties, such as "mission name", "mission duration", "mission" and "launch pad", for identified entity "Apollo 14", with "Astronauten" and "astronauts".

Our approach can also retrieve the partial set of appropriate results for more complex NL-Queries like "Gib mir alle Menschen, die in Wien geboren wurden und in Berlin gestorben sind"[11]. Therefore, we also report the performance of our system on the overall test dataset of 50 NL-Queries. The results are shown in Table 4. In this way, we can find the overall coverage of our approach on all types of NL-Queries.

Table 4. Evaluation on 50 German NL-Queries

	Average Precision	**Average Recall**	**F1**
Translation	0.217	0.24	0.228
Translation with ESA	0.34	0.386	0.361
CL-ESA	0.459	0.506	0.481

6 Conclusion and Future Work

This paper presented a novel approach to perform cross-lingual natural language querying over Linked Data that includes *entity search, linguistic analysis* and *cross-lingual semantic similarity and relatedness measure*. The approach was evaluated against 50 NL-Queries in German over DBpedia and achieved an average precision of 0.459, an average recall of 0.506 and F1 score of 0.361. However, on the NL-Queries that can be covered by this approach, the system achieved an average precision of 0.815, an average recall of 0.831 and a F1 score of 0.823. Our approach clearly shows that cross-lingual semantic similarity and relatedness measures outperform the automatic translation for cross-lingual NL-Querying over Linked Data.With this approach, cross-lingual information retrieval at document level can also be performed, if the documents are already marked up with the structured knowledge base. Therefore, we are planning to extend our approach for entity retrieval and associated documents, and evaluate the approach in the traditional information retrieval scenario.

Acknowledgments. This work has been funded in part by the European Union under Grant No. 258191 for the PROMISE project, as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Mika, P., Potter, T.: Metadata statistics for a large web corpus. In: WWW 2012 Workshop on Linked Data on the Web (2012)
2. Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D.R.B., Hiemstra, D., De Jong, F.: Wikitranslate: query translation for cross-lingual information retrieval using only wikipedia. In: Proceedings of the 9th CLEF (2009)

[11] "Give me all people that were born in Vienna and died in Berlin." in the English test dataset.

3. Jones, G., Fantino, F., Newman, E., Zhang, Y.: Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. In: CLIA 2008, p. 34 (2008)
4. Steinberger, R., Pouliquen, B., Ignat, C.: Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In: Proc. of the 4th Slovenian Language Technology Conf., Information Society (2004)
5. Freitas, A., Oliveira, J.G., O'Riain, S., Curry, E., Pereira da Silva, J.C.: Querying linked data using semantic relatedness: a vocabulary independent approach. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 40–51. Springer, Heidelberg (2011)
6. Unger, C., Bhmann, L., Lehmann, J., Ngomo, A.C.N., Gerber, D., Cimiano, P.: Sparql template based question answering. In: 21st International World Wide Web Conference, WWW 2012 (2012)
7. Yahya, M., Berberich, K., Elbassuoni, S., Ramanath, M., Tresp, V., Weikum, G.: Natural language questions for the web of data. In: EMNLP-CoNLL 2012 (2012)
8. Lu, C., Xu, Y., Geva, S.: Web-based query translation for english-chinese CLIR. In: Computational Linguistics and Chinese Language Processing (CLCLP), pp. 61–90 (2008)
9. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval, 209–230 (2001)
10. Littman, M., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using latent semantic indexing. In: Cross-Language Information Retrieval, ch. 5, pp. 51–62. Kluwer Academic Publishers (1998)
11. Zhang, D., Mei, Q., Zhai, C.: Cross-lingual latent topic extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 1128–1137. Association for Computational Linguistics, Stroudsburg (2010)
12. Sorg, P., Braun, M., Nicolay, D., Cimiano, P.: Cross-lingual information retrieval based on multiple indexes. In: Working Notes for the CLEF 2009 Workshop, Cross-Lingual Evaluation Forum, Corfu, Greece (2009)
13. Ferrández, Ó., Spurk, C., Kouylekov, M., Dornescu, I., Ferrández, S., Negri, M., Izquierdo, R., Tomás, D., Orasan, C., Neumann, G., Magnini, B., Vicedo, J.L.: The qall-me framework: A specifiable-domain multilingual question answering architecture. Web Semantics, 137–145 (2011)
14. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606–1611 (2007)
15. Sorg, P., Cimiano, P.: An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In: Horacek, H., Métais, E., Muñoz, R., Wolska, M. (eds.) NLDB 2009. LNCS, vol. 5723, pp. 36–48. Springer, Heidelberg (2010)