

# Learning a Cross-Lingual Semantic Representation of Relations Expressed in Text

Achim Rettinger, Artem Schumilin, Steffen Thoma, and Basil Ell

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

{rettinger, steffen.thoma, basil.ell}@kit.edu  
artem.schumilin@student.kit.edu

**Abstract.** Learning cross-lingual semantic representations of relations from textual data is useful for tasks like cross-lingual information retrieval and question answering. So far, research has been mainly focused on cross-lingual entity linking, which is confined to linking between phrases in a text document and their corresponding entities in a knowledge base but cannot link to relations. In this paper, we present an approach for inducing clusters of semantically related relations expressed in text, where relation clusters i) can be extracted from text of different languages, ii) are embedded in a semantic representation of the context, and iii) can be linked across languages to properties in a knowledge base. This is achieved by combining multi-lingual semantic role labeling (SRL) with cross-lingual entity linking followed by spectral clustering of the annotated SRL graphs. With our initial implementation we learned a cross-lingual lexicon of relation expressions from English and Spanish Wikipedia articles. To demonstrate its usefulness we apply it to cross-lingual question answering over linked data.

**Keywords:** Unsupervised Relation Extraction, Cross-lingual Relation Clustering, Relation Linking

## 1 Motivation

Due to the variability of natural language, a relation can be expressed in a wide variety of ways. When counting how often a certain pattern is used to express a relation (e.g. which movie is starring which actor), the distribution has a very long tail: frequently used patterns make up only a small fraction ; the majority of expressions use rare patterns (see Welty et al., [18]). While it would be possible to manually create patterns for a small set of languages, this would be a tedious task, results would not necessarily be correct, and coverage would most likely be far from optimal due to the size of the long tail. Thus, automatically extracting a set of syntactical variants of relations from text corpora would ease this task considerably.

However, there are numerous challenges associated to automating this task. It is essential to capture the context in which such a pattern applies. Typically, all of the information conveyed in a sentence is crucial to disambiguate the meaning of a relation expressed in text. Thus, a rich meaning representation is needed that goes beyond simple patterns consisting of named entity pairs and the string in-between them. Furthermore, semantically related relations need to be detected, grouped and linked to existing formalized knowledge. The latter is essential, if the meaning of the learned representations need to be related to human conceptualizations of knowledge, like questions answering over linked data. Finally, another dimension of complexity arises when we also consider the variability of natural language across different languages (e.g., English and Spanish). Then, finding patterns, aligning semantically related ones across languages, and linking them to one existing formal knowledge representations requires the learning of a cross-lingual semantic representation of relations expressed in text of different languages.

Unsupervised learning of distributional semantic representations from textual data has received increasing attention in recent years [10], since such representations have shown to be useful for solving tasks like document comparison, information retrieval and question answering. However, research has focused almost exclusively on the syntactic level and on single languages. At the same time, there has been progress in the area of cross-lingual entity disambiguation and linking, but this work is mostly confined to (named) entities and does not extend to other expressions in text, like the phrases indicating the relations between entities. What is missing so far is a representation that links linguistic variations of semantically related and contextualized textual elements across languages to their corresponding relation in a knowledge base.

In this paper, we present the first approach to unsupervised clustering of semantically related and cross-lingual relations expressed in text. This is achieved by combining multi-lingual semantic role labeling (SRL) with cross-lingual entity linking followed by spectral clustering of the resulting annotated SRL graphs. The resulting cross-lingual semantic representation of relations is, whenever possible, linked to English DBpedia properties, and enables e.g., to extend the schema with new properties, or to support cross-lingual question answering over linked data systems.

In our initial implementation we built a cross-lingual library of relation expressions from English and Spanish Wikipedia articles containing 25,000 SRL graphs with 2000 annotations to DBpedia entities. To demonstrate the usefulness of this novel language resource we show its performance on the Multi-lingual Question Answering over Linked Data challenge (QALD-4)<sup>1</sup>. Our results show that we can clearly outperform baseline approaches in respect to correctly linking (English) DBpedia properties in the SPARQL queries, specifically in a cross-lingual setting where the question to be answered is provided in Spanish.

---

<sup>1</sup> <http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=task1&q=4>

In summary, the main contributions of our proposed approach to extract, cluster and link contextualized relation expressions in text are the following:

- Relation expressions can be extracted from text of different languages and are not restricted to a predefined set of relations (as defined by DBpedia).
- Extracted expressions are embedded in a semantic graph, describing the context this expression appears in.
- Semantically related relation expressions and their associated context are disambiguated and clustered across languages.
- If existing, relation clusters are linked to their corresponding property in the English DBpedia.

In the remainder of this paper we first discuss related work, before introducing our approach to learning a cross-lingual semantic representation of grounded relations (Sec. 3 - 6). In Sec. 7 we evaluate our initial implementation on the QALD-4 benchmark and conclude in Sec. 8.

## 2 Related Work

Lewis and Steedman [7] present an approach to learning clusters of semantically equivalent English and French binary relations between referring expressions. Similar to us, a cluster is a language-independent semantic representation that can be applied to a variety of tasks such as translation, relation extraction, summarization, question answering, and information retrieval. The main difference is that we perform clustering on Semantic Role Label (SRL) graphs – thus operating on an abstract meaning representation - instead of binary syntactic relations. A meaning representation is more language-independent than a syntactic representation (like string patterns or dependency trees) since it abstracts from grammatical variations of different languages. This facilitates the learning of cross-lingual and language-independent semantic representations. This basic difference applies to almost all of the remaining approaches listed in this section, like Lin and Pantel (DIRT, [8]), who learn textual inference rules such as ("X wrote Y", "X is the author of Y") from dependency-parsed sentences by building groups of similar dependency paths.

An additional difference of related approaches like [17,11,16,5,14] is their dependency on preexisting knowledge base properties. In contrast, our approach does not start from a predefined set of knowledge base property for which we learn textual representations, but instead derives clusters of textual expressions via Semantic Role Labeling first for which we then try to find a corresponding relation in the KB. Thus, our approach is not confined to finding relations preexisting in a knowledge base. Newly identified relations could even be used for extending the ontology. This, however, would be contribution to ontology learning and is out of the scope of this paper. The approaches restricted to preexisting KB relations (and shallow parsing) are discussed in more detail now. Walter et al. (M-ATOLL, [17]) learn dependency paths as natural language expressions for KB relations. They begin with a relation from DBpedia, retrieve

triples for this relation and search within a text corpus for sentences where the two arguments of the relation can be found within a sentence. The sentence is dependency-parsed and, given a set of six dependency patterns, a pattern matches the dependency tree. Mahendra et al. ([11]) learn textual expressions of DBpedia relations from Wikipedia articles. Given a relation, triples are retrieved and sentences are identified where the two arguments of the relation can be found within a sentence. The longest common substring between the entities in sentences collected for a relation is learned as the relation’s textual expression. Vila et al. (WRPA, [16]) learn English and Spanish paraphrases from Wikipedia for four pre-specified relations. Textual triples are derived using data from an article’s infobox and its name. The string between the arguments of a relation within a sentence is extracted and generalized and regular expressions are created. Gerber and Ngonga Ngomo (BOA, [5]) language-independently learn textual expressions of DBpedia relations from Wikipedia by regarding the strings between a relation’s arguments within sentences. Nakashole et al. (PATY, [14]) learn textual expressions of KB relations from dependency-parsed or POS-tagged English sentences. Textual expressions are sequences of words, POS-tags, wildcards, and ontological types.

In contrast to the work just mentioned, there are a few approaches that leverage a semantic representation. Grounded Unsupervised Semantic Parsing by Poon (GUSP, see [15]) translates natural-language questions to database queries via a learned probabilistic grammar. However, GUSP is not cross-lingual. Similarly, Exner and Nugues [4] learn mappings from PropBank to DBpedia based on Semantic Role Labeling. Relations in Wikipedia articles are detected via SRL, named entities are identified and linked to DBpedia and use these links to ground PropBank relations to DBpedia. Again, this is not cross-lingual.

To the best of our knowledge, our approach is the only one that i) extracts potentially novel relations and ii) where possible, links to preexisting relation in a KB and iii) does this across languages by exploiting a language-independent semantic representation rather than a syntactic one.

### 3 A Pipeline for Learning a Cross-lingual Semantic Representations of Grounded Relations

Our pipeline, as shown in Figure 1, consists of three major stages.

In the first stage (see Sec. 4), the multi-lingual text documents are transformed and processed by Semantic Role Labeling (SRL). In our evaluation we use Wikipedia articles as input data, but any text that produces valid SRL graphs is feasible. Please note, to construct a cross-lingual representation a multi-lingual comparable corpus covering similar topics is advisable. However, there is no need for a aligned or parallel corpus. SRL produces semantic graphs of frames with predicates and associated semantic role-argument pairs. In parallel, we apply cross-lingual entity linking to the same text documents. This detects entity mentions in multi-lingual text and annotates the corresponding mention strings with the entity URI originating exclusively from the English

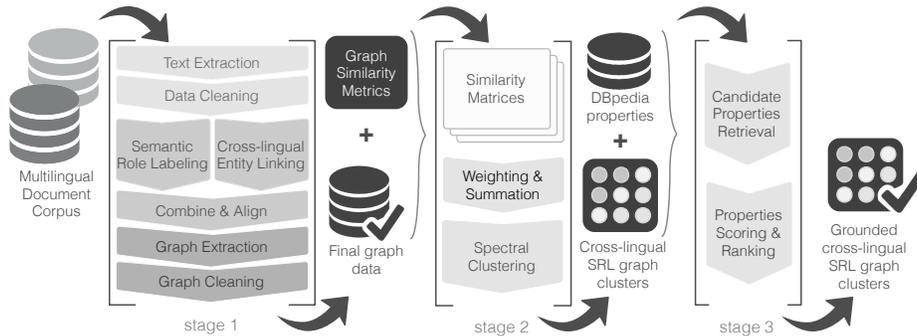


Fig. 1: Schematic summary of the processing pipeline.

DBpedia. After that, we combine and align the output of both, SRL and entity linking in order to extract a cross-lingual SRL graphs. The only remaining language-dependent elements in a cross-lingual SRL graph are the predicate nodes.

The next stage performs relational learning of cross-lingual clusters (Sec. 5) on the previously acquired annotated SRL graphs. The similarity metrics that we define in Section 5.1 are central to this stage of the pipeline.

In the subsequent third stage, the obtained clusters are linked to DBpedia properties. Section 6 describes this procedure in greater detail. As a result we get cross-lingual clusters of annotated SRL graphs, i.e. textual relation expressions, augmented with a ranked set of DBpedia properties. Ultimately, these grounded clusters of relation expressions are evaluated in the task of property linking on multi-lingual questions of the QALD-4 dataset.

## 4 Extracting and Annotating SRL Graphs

Multi-lingual Semantic Role Labeling is performed on the input text independently for every language. SRL is accomplished by means of shallow and deep linguistic processing as described in [9]. The result of this processing step is a semantic graph consisting of semantic frames with predicates and their arguments. Each semantic frame is represented as a tree with the predicate as the root and its arguments as the leaf nodes. The edges are given by the semantic roles of the predicate arguments (cmp. Fig. 2).

SRL graphs are directed, node and edge labelled graphs describing the content of a whole document. Several predicates appear in one graph, so one sub-tree per predicate is extracted for clustering (the predicate being the root of the tree), resulting in a few trees per sentence and many trees per document. Trees from one document contain partially duplicated information. Formally, an SRL graph is a set of triples  $t = (p, r, v)$  where the predicate  $p$  belongs to a set of SRL predicates ( $p \in P_{SRL}$ ), the role  $r$  belongs to a set of SRL roles ( $r \in R_{SRL}$ ), and  $v$  is either a string value or an SRL predicate ( $v \in P_{SRL} \cup String$ ). We consider a frame as valid, if it has at least two non-frame arguments. Such a

Only a few cults were banned by the Roman authorities...

```
<frame displayName="ban.01" ID="F541" sentenceID="57" tokenID="57.6" >
  <argument displayName="cult" role="A1:Theme" id="W544" />
  <argument displayName="imperial_roman" role="A0:Agent" id="E1" />
  <argument displayName="be.00" role="AM-ADV" frame="true" id="F542" />
  <descriptions>
    <description URI="00796392-v" displayName="ban" knowledgeBase="WordNet-3.0" />
  </descriptions>
</frame>

<DetectedTopic URL="http://dbpedia.org/resource/Cult_(religious_practice)" mention="cults"
  displayName="Cult (religious practice)" from="7064" to="7069" weight="0.01" \>
<DetectedTopic URL="http://dbpedia.org/resource/Roman_Empire" mention="Roman authorities"
  displayName="Roman Empire" from="7089" to="7106" weight="0.393" \>
```

Fig. 2: Example sentence with corresponding partial XML outputs produced by SRL (*frame* element) and the cross-lingual entity linking tool (*DetectedTopic* elements).

constraint reduces the number of usable frames, which, in turn is compensated by the large amount of the raw textual data.

The example in Fig. 2 demonstrates the operation of the SRL pipeline, beginning with an example sentence for which the semantic frame is obtained. To achieve cross-lingual SRL graphs role labels of non-English SRL graphs are mapped to their corresponding English role labels. Whenever possible SRL predicates from all languages are linked to English wordnet synsets. That's not always possible since not every phrase of a predicate in an extracted SRL graph is mentioned in WordNet, specifically for non-English languages.

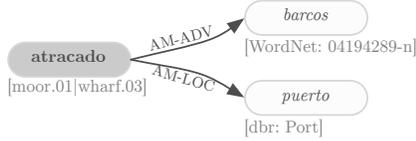
The next step towards generating cross-lingual SRL graphs is cross-lingual entity linking to the English DBpedia. This language-independent representation of the predicate arguments provides additional cross-lingual context for the subsequent predicate cluster analysis.

We treat this step as a replaceable black-box component by using the approach described in [19]. [19] relies on linkage information in different Wikipedia language versions (language links, hyper links, disambiguation pages, ...) plus a statistical cross-lingual text comparison function, trained on a comparable corpora. The cross-lingual nature of our analysis is achieved by mapping text mentions in both languages to the English-language DBpedia URIs. The bottom part of Fig. 2 is a sample of the annotation output for the above example sentence. Annotations that correspond to SRL arguments are enclosed in *URL* attributes of *DetectedTopic* elements.

The intermediate results of both, the SRL and annotation steps finally need to be combined in order to extract the actual graphs. Figure 3 contains an example of four sentences along with the extracted cross-lingual SRL graphs from English and Spanish sentences. The graph vertices show the SRL predicate and argument mention strings along with DBpedia URIs (*dbr* namespace <http://dbpedia.org/resource/>) and Wordnet-IDs. Edge labels specify the semantic role. Obviously, the graphs on the top and on the bottom are more similar to

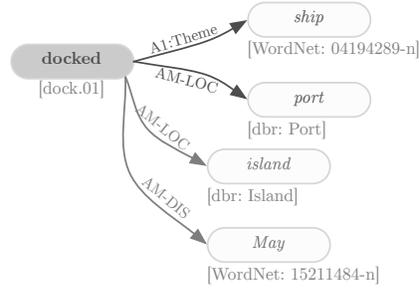
Spanish sentence 1:

En mayo de 1937 el Deutschland estaba **atracado** en el *puerto* de Palma, en Mallorca, junto con otros *barcos* de guerra neutrales de las armadas británica e italiana.



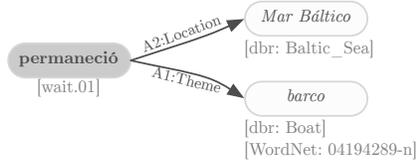
English sentence 2:

In *May* 1937, the *ship* was **docked** in the *port* of Palma on the *island* of Majorca, along with several other neutral warships, including vessels from the British and Italian navies.



Spanish sentence 3:

Los problemas en sus motores obligaron a una serie de reparaciones que culminaron en una revisión completa a fines de 1943, tras lo que el *barco* **permaneció** en el *Mar Báltico*.



English sentence 4:

Engine problems forced a series of repairs culminating in a complete overhaul at the end of 1943, after which the *ship* **remained** in the *Baltic*.

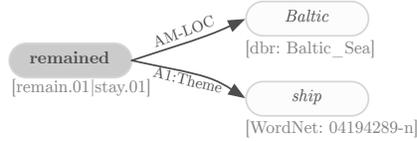


Fig. 3: Cross-lingual SRL graphs extracted from English and Spanish sentences.

each other compared to the graphs on the level and right, respectively. Thus, cross-lingual SRL graphs are similar regarding the content, not the language.

## 5 Learning a Cross-Lingual Semantic Representation of Relation Expressions

For the purpose of clustering a set of cross-lingual SRL graphs we introduce a set of metrics specifying a semantic distance of SRL graphs (see Sec. 5.1). Section 5.2 discusses the spectral clustering algorithm.

### 5.1 Constructing Similarity Matrices of Annotated SRL Graphs

Goal of this step is to construct a similarity matrix, specifying the pair-wise similarity of all SRL graphs. We tried three different graph-similarity metrics  $m_1, m_2, m_3$ .

Formally, a cross-lingual SRL graph is an SRL graph where  $v$  is either a string value, an SRL predicate, or a unique identifier ( $v \in P_{SRL} \cup String \cup \mathcal{U}$ ).  $g(p)$  denotes the graph with predicate  $p$  as the root SRL predicate.  $m_1 : G \times G \rightarrow \{1; 0\}$  compares the SRL graphs' root predicates according to their names, e.g. *exist.01* vs. *meet.02*:

$$m_1(g_i, g_j) := \begin{cases} 1, & p(g_i) = p(g_j) \\ 0, & \text{else} \end{cases} \quad (1)$$

$m_2 : G \times G \rightarrow [1; 0]$  compares two SRL graphs' root predicates according to their annotated role values:

$$m_2(g_i, g_j) := \frac{|A(g_i) \cap A(g_j)|}{|A(g_i) \cup A(g_j)|} \quad (2)$$

where  $A(g_k) := \{v \mid \exists r \in R_{SRL} : (p(g_k), r, v) \in g_k \wedge v \in \mathcal{U}\}$ .

$m_3 : G \times G \rightarrow [1; 0]$  compares two SRL graphs' root predicates according to their role labels:

$$m_3(g_i, g_j) := \frac{|B(g_i) \cap B(g_j)|}{|B(g_i) \cup B(g_j)|} \quad (3)$$

where  $B(g_k) := \{r \mid \exists v \in P_{SRL} \cup String \cup \mathcal{U} : (p(g_k), r, v) \in g_k\}$ .

Now, given the set of cross-lingual SRL graphs  $\{g_1, \dots, g_n\}$  and given the three SRL predicate similarity metrics, we can construct three SRL predicate similarity matrices. Each SRL predicate similarity metric is applied for pairwise comparison of two (annotated) SRL graphs' root predicates. The root predicate  $p$  of an (annotated) SRL graph  $g$ , denoted by  $p(g)$ , is the predicate for which no triple  $(p_2, r, p) \in g$  exists with  $p \neq p_2$ .  $G$  denotes the set of all SRL graphs.

Based on a separate evaluation of each metric we introduce a combined similarity metric as a weighted sum of the three single metrics.

## 5.2 Spectral Clustering of Annotated SRL Graphs

Spectral Clustering uses the spectrum of a matrix derived from distances between different instances. Using the spectrum of a matrix has been successfully used in many computer vision applications [12] and is also applicable for similarity matrices. As input a similarity matrix  $S$  derived from one metric or a weighted combination of several metrics is given. As a first step the Laplacian matrix  $L$  is built by subtracting the similarity matrix  $S$  from the diagonal matrix  $D$  which contains the sum of each row on the diagonal (respectively column since  $S$  is symmetric) (Eq. 4).

$$L_{ij} = D_{ij} - S_{ij} = \begin{cases} \sum_m S_{im} - S_{ij} = \sum_m S_{mj} - S_{ij} & \text{if } i = j \\ -S_{ij} & \text{otherwise} \end{cases} \quad (4)$$

For building  $k$  clusters, the second up to the  $k + 1$  smallest eigenvalue and corresponding eigenvector of the Laplacian matrix are calculated. Afterwards the actual clustering starts with running the  $k$ -means algorithm on the eigenvectors which finally results in a clustering for the instances of  $S$ .

To enforce the learning of cross-lingual clusters, we introduce the weighting matrix  $W$  which is used to weight the mono- and cross-lingual relations in the similarity matrix  $S$  (Eq. 5). While setting the monolingual weight  $w_{monolingual}$  to zero, forces the construction of only cross-lingual clusters, we received better

results by setting  $w_{monolingual} > 0$ . This can be intuitively understood as we get more clean clusters when we don't force cross-lingual relations into every cluster, as there is no guarantee that a matching cross-lingual relation even exists. Finally the weighted matrix  $S^*$ , the result of the product  $W$  and  $S$  (Eq. 6), is given as input to the previously described spectral clustering algorithm.

$$W_{ij} = \begin{cases} w_{monolingual} & \text{if } i \text{ and } j \text{ are monolingual} \\ 1 & \text{if } i \text{ and } j \text{ are crosslingual} \end{cases} \quad (5)$$

$$S_{ij}^* = W_{ij} \cdot S_{ij} \quad (6)$$

## 6 Linking Annotated SRL Graph Clusters to DBpedia Properties

In order to find potential links of the obtained clusters to DBpedia properties, we exploit the SRL graphs' argument structure as well as the DBpedia entity URIs provided by cross-lingual entity linking. The origin of possible candidates is limited to the DBpedia ontology<sup>2</sup> and infobox<sup>3</sup> properties.

*Acquisition of Candidate Properties* For a given annotated SRL graph we retrieve a list of candidate properties by querying DBpedia for the in- and outbound properties associated with its arguments' entities. Consequently, the candidate properties of an entire predicate cluster are determined by the union of the individual graphs' candidate lists. Several specific properties, such as the Wikipedia-related structural properties (e.g. wikiPageID, wikiPageRevisionID etc.) are excluded from the candidate list.

*Scoring of Candidate Properties* After the construction of the candidate list, the contained properties are scored. The purpose behind this is to determine a ranking of properties by their relevance with respect to a given cluster. In principle, several different scoring approaches are applicable to the underlying problem. For example, a relative normalized frequency score of property  $p_i$  w.r.t. cluster  $C_j$  calculated as

$$S^{rf}(p_i, C_j) = \frac{\text{relative frequency of } p_i \text{ in } C_j}{\text{relative frequency of } p_i \text{ over all clusters}}$$

is appropriate to reflect the importance as well as the exclusiveness of property  $i$  for cluster  $j$ . However, our experiments determined the absolute frequency score of a property within a cluster to be the best performing measure.

Alg. 1 shows the structure of the complete grounding algorithm in a simplified form. This algorithm is similar to the approach by Exner and Nugues [4].

<sup>2</sup> URI namespace <http://dbpedia.org/ontology/>

<sup>3</sup> URI namespace: <http://dbpedia.org/property/>

---

**Algorithm 1** Algorithm that computes a ranked set of DBpedia properties for a given relation cluster.

---

**Input:** SRL graph cluster  $c$

$result \leftarrow \emptyset$

**for all**  $p \in \{p^{KB} \mid \exists g \in c : \exists(p^{SRL}, r, e) \in g : (\exists o : (e, p^{KB}, o) \in KB \vee \exists s : (s, p^{KB}, e) \in KB)\}$  **do**

$result \leftarrow result \cup (p, |\{(s, p, o) \in KB \mid \exists g \in c : (p^{SRL}, r, e) \in g : e \in \mathcal{R} \wedge (s = e \vee o = e)\}|)$

**end for**

**Return:** result

---

## 7 Evaluation on Cross-lingual Relation Linking for Question Answering over Linked Data

We make use of the evaluation data set provided by the Multi-lingual Question Answering over Linked Data challenge (task 1 of QALD-4). The data set contains 200 questions (12 out of 200 are out-of-scope w.r.t DBpedia knowledge base) in multiple languages as well as corresponding gold-standard SPARQL queries against DBpedia.

To evaluate the quality of our results, we conducted property linking experiments. We deliberately concentrate on the sub-task of property linking to avoid distortion of the performance by various pre- and post-processing steps of a full QA-system. Linking the properties necessary for constructing the SPARQL query constitutes an important step of a question answering system such as QAKiS [1], SemSearch [6], ORAKEL [2], FREyA [3], and TcruziKB [13] which generate SPARQL queries based on user input.

### 7.1 Linking Properties in the QALD challenge

First, we generated compatible data representation from the QALD-4 question sentences by sending them through stage 1 of our processing pipeline (see Sec. 3). Hereby we obtained cross-lingual SRL graphs for English and Spanish questions.

Next, using our similarity metrics and the previously learned grounded clusters, we classified each individual SRL graph of the questions set and determined its target cluster. Consequently, each SRL graph of the questions set was assigned DBpedia properties according to the groundings of its associated target cluster. This way, for each question, our approach linked properties, which were finally evaluated against the gold-standard properties of the QALD-4 training dataset.

### 7.2 Data Set and Baselines

We employed Wikipedia as the source of multi-lingual text documents in the English (EN, Wikipedia dump version 2013.04.03) and Spanish (ES, Wikipedia dump version 2012.05.15) language. Over 23,000,000 cross-lingual

annotated SRL graphs were extracted from more than 300,000 pairs of language link-connected English and Spanish Wikipedia articles.

In order to get an initial assessment of our approach we conducted our experiments on two samples of the original data. Table 1 provides an overview of the key dataset statistics. Dataset 1 consists of a random sample of long Wikipedia article pairs, which together sum up to approximately 25,000 SRL graph instances. The second sample with a similar number of graphs was derived from randomly selected short article pairs in order to provide a wider coverage of different topics and corresponding DBpedia entities.

	Dataset 1: "long articles"		Dataset 2: "short articles"	
	English	Spanish	English	Spanish
# documents	29	29	1,063	1,063
# extracted graphs	10,421	14,864	13,009	12,402
# mentioned DBpedia entities	2,065		13,870	
# unique DBpedia entities	1,379		6,300	

Table 1: Key statistics of the data sets used for our experiments.

**Baseline 1: String Similarity-based Property Linking** This first naïve baseline links properties based on string similarity between the question tokens and DBpedia property labels. Given a question from the QALD-4 training dataset, we firstly obtain the question tokens using the Penn treebank-trained tokenizer. In the next step, each token is assigned the one DBpedia property with the highest string similarity between its label and the token string. String similarity is measured by means of the normalized Damerau-Levenshtein distance. For each token, the one property with the highest label similarity enters the candidate set. Finally, the identified candidate properties are evaluated against the QALD-4 gold-standard properties. Because the vast majority of property labels are of English origin, we could not apply this baseline to Spanish QALD-4 data.

**Baseline 2: Entity-based Property Linking** Baseline 2 takes a more sophisticated approach to finding good candidate properties. For this baseline, we first use the set of entities associated with a given question for linking of candidate properties exactly the same way as we perform grounding of cross-lingual SRL graph clusters (Sec. 5.1). In the next step, the list of candidate properties is pruned by thresholding the normalized Damerau-Levenshtein similarity of their labels to the question tokens. Again, this will have negative effect on the performance for Spanish-language questions for the same reasons as discussed in 7.2. We report results for two variations of this baseline, which differ in the mode of entity retrieval for a given question: In the first case, entities are collected from the cross-lingual annotated SRL graphs, while in the second case we obtain the entities directly from the output of the entity linking tool.

### 7.3 Evaluation Results

**Baseline 1: Results** A naïve selection of candidate properties based solely on string similarity between the question tokens and property labels shows poor overall performance on the English-language QALD-4 questions:

precision: 2.15%      recall: 10.68%      F1-measure: 3.58%

As discussed in 7.2, this baseline is limited to English-language questions.

**Baseline 2: Results** The top part of Table 2 shows the performance of Baseline 2 in the case *without* SRL graph extraction.

		string similarity threshold					
		0.4	0.5	0.6	0.7	0.8	0.9
WITHOUT SRL	precision EN [%]	2.2	5.0	11.3	19.3	21.9	21.6
	precision ES [%]	0.7	1.9	5.0	6.3	12.5	21.4
	F1-measure EN [%]	4.1	8.4	15.7	22.6	23.2	22.3
	F1-measure ES [%]	1.4	2.9	6.0	6.8	14.3	22.0
WITH SRL	precision EN [%]	3.2	6.7	16.8	24.3	23.5	22.5
	precision ES [%]	0.7	1.9	5.6	3.2	10.0	0.0
	F1-measure EN [%]	5.4	9.7	19.2	26.5	24.5	22.5
	F1-measure ES [%]	1.2	2.5	6.2	3.1	10.5	0.0

Table 2: Performance of Baseline 2 without and with SRL graph extraction.

Due to the cross-lingual nature of property linking through our grounding algorithm, there is a clear performance increase for Spanish-language questions. It is also notable that the behaviour of the performance measure is consistent over all string similarity thresholds for both languages. The bottom part of Table 2 shows Baseline 2 results with SRL graph extraction. Here, we see a small but consistent performance increase for the English language over Baseline 2 without SRL. This observation supports our assumption that the inclusion of the semantic structure of annotated arguments as provided by Semantic Role Labeling does improve performance.

**Results with Grounded Cross-lingual SRL Graph Clusters** The evaluation of our approach was conducted on the previously described (Tab. 1) experimental datasets and a variety of different clustering configurations with respect to different similarity matrices as well as different internal parameter sets of the spectral clustering algorithm.

Table 3 reports the results of several top performing configurations. It is notable that across languages and different parameter sets, the completely cross-lingual, entity-focused metric  $m_2$  outperforms the other configurations, which supports the basic idea of our approach. In addition to this, we observe a

lang.	clustering configuration				performance [%]		
	metric	#clusters	#eigenvectors	$w_{monolingual}$	precision	recall	F1
ES	m2	500	100	0.0	30.19	28.57	29.36
ES	m2	200	100	0.0	30.05	28.44	29.22
ES	m2	100	50	0.0	30.05	28.19	29.09
ES	m2	200	50	0.0	29.77	28.19	28.96
EN	m2	200	50	0.0	29.52	27.24	28.33
EN	m2	100	50	0.0	29.44	27.09	28.22
EN	m2	200	100	0.0	29.13	26.91	27.97
EN	m2	10	50	0.0	28.99	26.74	27.82

Table 3: Best performing grounded clusters configurations for QALD-4 questions.

lang.	clustering configuration				performance [%]		
	dataset	# clusters	# eigenvectors	$w_{monolingual}$	precision	recall	F1
EN	2 (short)	200	100	0.0	27.09	26.25	26.67
EN	2 (short)	200	50	0.0	24.12	23.85	23.98
ES	2 (short)	200	100	0.0	28.70	27.47	28.07
ES	2 (short)	200	50	0.0	27.68	26.50	27.07
EN	1 (long)	200	100	0.0	21.30	21.00	21.15
EN	1 (long)	200	100	0.0	20.38	20.19	20.28
ES	1 (long)	200	50	0.0	21.33	20.87	21.10
ES	1 (long)	200	50	0.0	18.98	18.64	18.81

Table 4: Best performing results for "short articles" vs "long articles".

consistent improvement over our baselines for English, and even more so for the Spanish language.

To investigate the effect of input data and parameter choice on the quality of results, we conducted further experiments, which involved grounded clusters computed on a weighted sum of all metrics with cross-lingual constraints. In particular, we demonstrate the effect of the short- versus long-articles dataset, i.e. the impact of more diverse input data. Table 4 shows results of this comparison. Obviously, shorter and more concise articles seem to produce SRL graphs with more meaningful clusters. It would be interesting to evaluate whether co-reference resolution would improve the performance for longer articles.

Another aspect of interest is the effect of the number of Eigenvectors within the spectral clustering algorithm. This parameter greatly increases the computational resources needed to compute the clustering. But our experimental results also clearly show an advantage of a high number of Eigenvectors (Tab. 5).

Both experiments revealed that more input data as well as higher-dimensional clustering has the potential to further improve the performance of our approach. Another incentive for scaling those dimensions is to cover the long tail of relation expressions. Still, we would argue that this limited evaluation clearly demonstrates the benefits of our approach, since we outperform Baseline 2 by about 6% and Baseline 2 is comparable to what is used in most of the related work. That shows a big potential to improve those QA systems.

lang.	clustering configuration				performance [%]		
	dataset	#clusters	#eigenvectors	$w_{monolingual}$	precision	recall	F1
EN	2 (short)	500	500	0.5	27.65	27.15	27.04
EN	2 (short)	200	200	0.5	27.23	26.87	27.05
ES	2 (short)	200	500	0.5	29.09	27.35	28.19
ES	2 (short)	200	300	0.5	29.09	27.35	28.19
EN	2 (short)	200	50	0.5	25.00	24.56	24.77
EN	2 (short)	500	50	0.5	21.58	21.49	21.53
ES	2 (short)	200	50	0.5	18.02	17.94	17.98
ES	2 (short)	500	50	0.5	13.24	13.24	13.24

Table 5: Best performing results in respect to number of eigenvectors.

## 8 Conclusion and Future Work

This paper introduces an approach to unsupervised learning of a cross-lingual semantic representation of relations expressed in text. To the best of our knowledge this is the first meaning representation induced from text that is i) cross-lingual, ii) builds on semantic instead of shallow syntactic features, and iii) generalizes over relation expressions. The resulting clusters of semantically related relation graphs can be linked to DBpedia properties and thus support tasks like question answering over linked data. Our results show that we can clearly outperform baseline approaches on the sub-task of property linking.

Directions for future work include, learning the semantic representation from more documents. Our current implementation serves as a strong proof-of-concept, but does not yet cover the long tail of relation expressions sufficiently. Including all Wikipedia articles resulting in millions of graphs is merely an engineering challenge, only the clustering step would need to be adjusted. In addition, we would like to assess the potential of our approach to discover novel relation-types (and their instantiations) to the knowledge base.

**Acknowledgments.** The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

## References

1. E. Cabrio, J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli, and F. G. Qakis: an open domain qa system based on relational patterns. In *In Proc. of the 11th International Semantic Web Conference (ISWC 2012), demo paper*, 2012.
2. P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer. Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system. *Data & Knowledge Engineering*, 65(2):325–354, 2008.
3. D. Damjanovic, M. Agatonovic, and H. Cunningham. FREyA: An interactive way of querying Linked Data using natural language. In *Proceedings of the 8th International Conference on The Semantic Web, ESWC ’11*, pages 125–138, 2012.

4. P. Exner and P. Nugues. Ontology matching: from propbank to dbpedia. In *SLTC 2012, The Fourth Swedish Language Technology Conference*, pages 67–68, 2012.
5. D. Gerber and A.-C. N. Ngomo. Extracting Multilingual Natural-language Patterns for RDF Predicates. *EKAW'12*, pages 87–96, 2012.
6. Y. Lei, V. Uren, and E. Motta. SemSearch: A Search Engine for the Semantic Web. In *Managing Knowledge in a World of Networks*, pages 238–245. 2006.
7. M. Lewis and M. Steedman. Unsupervised induction of cross-lingual semantic relations. In *EMNLP*, pages 681–692, 2013.
8. D. Lin and P. Pantel. Dirt - discovery of inference rules from text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2001, pages 323–328. ACM, 2001.
9. X. Lluís, X. Carreras, and L. Màrquez. Joint arc-factored parsing of syntactic and semantic dependencies. *Transactions of the Association of Computational Linguistics – Volume 1*, pages 219–230, 2013.
10. P. S. Madhyastha, X. Carreras Pérez, and A. Quattoni. Learning task-specific bilingual embeddings. In *Proceedings of COLING-2014*, 2014.
11. R. Mahendra, L. Wanzare, R. Bernardi, A. Lavelli, and B. Magnini. Acquiring relational patterns from wikipedia: A case study. In *Proceedings of the 5th Language and Technology Conference*, 2011.
12. J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *Int. J. Comput. Vision*, 43(1):7–27, June 2001.
13. P. N. Mendes, B. McKnight, A. P. Sheth, and J. C. Kissinger. TcruziKB: Enabling Complex Queries for Genomic Data Exploration. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing, ICSC '08*, pages 432–439, 2008.
14. N. Nakashole, G. Weikum, and F. Suchanek. Patty: A taxonomy of relational patterns with semantic types. *EMNLP-CoNLL '12*, pages 1135–1145, 2012.
15. H. Poon. Grounded unsupervised semantic parsing. In *ACL (1)*, pages 933–943. Citeseer, 2013.
16. M. Vila, H. Rodríguez, and A. M. Mart'i. Wrpa: A system for relational paraphrase acquisition from wikipedia. *Procesamiento del lenguaje natural*, (45):11–19, 2010.
17. S. Walter, C. Unger, and P. Cimiano. M-ATOLL: A Framework for the Lexicalization of Ontologies in Multiple Languages. In *Proceedings of the 13th International Conference on The Semantic Web, ISWC 2014*, pages 472–486, 2014.
18. C. Welty, J. Fan, D. Gondek, and A. Schlaikjer. Large scale relation detection. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, FAM-LbR '10*, pages 24–33, 2010.
19. L. Zhang and A. Rettinger. X-lisa: Cross-lingual semantic annotation. *PVLDB*, 7(13):1693–1696, 2014.