Language-agnostic Relation Extraction from Wikipedia Abstracts

Nicolas Heist and Heiko Paulheim

Data and Web Science Group, University of Mannheim, Germany

Abstract. Large-scale knowledge graphs, such as DBpedia, Wikidata, or YAGO, can be enhanced by relation extraction from text, using the data in the knowledge graph as training data, i.e., using *distant supervision.* While most existing approaches use language-specific methods (usually for English), we present a language-agnostic approach that exploits background knowledge from the graph instead of language-specific techniques and builds machine learning models only from language-independent features. We demonstrate the extraction of relations from Wikipedia abstracts, using the twelve largest language editions of Wikipedia. From those, we can extract 1.6M new relations in DBpedia at a level of precision of 95%, using a RandomForest classifier trained only on language-independent features. Furthermore, we show an exemplary geographical breakdown of the information extracted.

1 Introduction

Large-scale knowledge graphs, like DBpedia [16], Freebase [3], Wikidata [30], or YAGO [17], are usually built using heuristic extraction methods, e.g., from Wikipedia infoboxes, by exploiting crowd-sourcing processes, or both. These approaches can help creating large-scale public cross-domain knowledge graphs, but are prone both to errors as well as incompleteness. Therefore, over the last years, various methods for refining those knowledge graphs have been developed [22]. For filling missing relations (e.g., the missing birthplace of a person), *relation extraction* methods are proposed. Those can be applied to fill in relations for entities derived from Wikipedia pages without or with only sparsely filled infoboxes.

Most methods for relation extraction work on text and thus usually have at least one component which is explicitly specific for the language at hand (e.g., stemming, POS tagging, dependency parsing), like, e.g., [10, 27, 35], or implicitly exploits some characteristics of that language [2]. Thus, adapting those methods to work with texts in different natural languages is usally not a straight forward process.

In this paper, we propose a language-agnostic approach. Instead of knowledge about the language, we take background knowledge from the DBpedia knowledge graph into account. With that, we try to discover certain patterns in how Wikipedia abstracts are written. For example, in many cases, any genre mentioned in the abstract about a band is usually a genre of that band, the first city mentioned in an abstract about a person is that person's birthplace, and so on. In that case, the linguistic assumptions that we make about a language at hand are quite minimal. In fact, we only assume that for each language edition of Wikipedia, there are certain ways to structure an abstract of a given type of entity, in terms of what aspect is mentioned where (e.g., the birth place is the first place mentioned when talking about a person). Thus, the approach can be considered as language-independent (see [2] for an in-depth discussion).

The choice for Wikipedia abstracts as a corpus mitigates one of the common sources of errors in the relation extraction process, i.e., the entity linking. Since Wikipedia articles can be unambiguously related to an instance in a knowledge base, and Wikipedia page links contained in Wikipedia abstracts are mostly free from noise, the corpus at hand can be directly exploited for relation extraction without the need for an upfront potentially noisy entity linking step.

By applying the exact same pipeline without any modifications to the twelve largest languages of Wikipedia, which encompass languages from different language families, we demonstrate that such patterns can be extracted from Wikipedia abstracts in arbitrary languages. We show that it is possible to extract valuable information by combining the information extracted from different languages.

The rest of this paper is structured as follows. In section 2, we review related work. We introduce our approach in section 3, and discuss various experiments in section 4. We conclude with a summary and an outlook on future work.

2 Related Work

Various approaches have been proposed for relation extraction from text, in particular from Wikipedia. In this paper, we particularly deal with *closed* relation extraction, i.e., extracting new instantiations for relations that are defined a priori (by considering the schema of the knowledge graph at hand, or the set of relations contained therein).

Using the categorization introduced in [22], the approach proposed in this paper is an *external* one, as it uses Wikipedia as an external resource in addition to the knowledge graph itself. While internal approaches for relation prediction in knowledge graphs exist as well, using, e.g., association rule mining, tensor factorization, or graph embeddings, we restrict ourselves to comparing the proposed approach to other external approaches.

Most of the approaches in the literature make more or less heavy use of language-specific techniques. Distant supervision is proposed by [19] as a means to relation extraction for Freebase from Wikipedia texts. The approach uses a mixture of lexical and syntactic features, where the latter are highly language-specific. A similar approach is proposed for DBpedia in [1]. Like the Freebase-centric approach, it uses quite a few language-specific techniques, such as POS tagging and lemmatization. While those two approaches use Wikipedia as a corpus, [13] compare that corpus to a corpus of news texts, showing that the usage of Wikipedia leads to higher quality results.

[20] introduce an approach for mining relations from Wikipedia articles which exploits similarities of dependency trees for extracting new relation instances. In [34], the similarity of dependency trees is also exploited for clustering pairs of concepts with similar dependency trees. The construction of those dependency trees is highly language specific, and consequently, both approaches are evaluated on the English Wikipedia only.

An approach closely related to the one discussed in this paper is *iPopulator* [15], which uses Conditional Random Fields to extract patterns for infobox values in Wikipedia abstracts. Similarly, *Kylin* [33] uses Conditional Random Fields to extract relations from Wikipedia articles and general Web pages. Similarly to the approach proposed in this paper, *PORE* [31] uses information on neighboring entities in a sentence to train a support vector machine classifier for the extraction of four different relations. The papers only report results for English language texts.

Truly language-agnostic approaches are scarce. In [8], a multi-lingual approach for open relation extraction is introduced, which uses Google translate to produce English language translations of the corpus texts in a preprocessing step, and hence exploits externalized linguistic knowledge. In the recent past, some approaches based on deep learning have been proposed which are reported to or would in theory also work on multi-lingual text [21, 29, 36, 37]. They have the advantages that (a) they can compensate for shortcomings in the entity linking step when using arbitrary text and (b) that explicit linguistic feature engineering is replaced by implicit feature construction in deep neural networks. In contrast to those works, we work with a specific set of texts, i.e., Wikipedia abstracts. Here, we can assume that the entity linking is mostly free from noise (albeit not complete), and directly exploit knowledge from the knowledge graph at hand, i.e., in our case, DBpedia.

In contrast to most of those works, the approach discussed in this paper works on Wikipedia abstracts in *arbitrary* languages, which we demonstrate in an evaluation using the twelve largest language editions of Wikipedia. While, to the best of our knowledge, most of the approaches discussed above are only evaluated on one or at maximum two languages, this is the first approach to be evaluated on a larger variety of languages.

3 Approach

Our aim is to identify and exploit typical patterns in Wikipedia abstracts. As a running example, we use the *genre* relation which may hold between a music artist and a music genre. Fig. 1 depicts this example with both an English and a French Wikipedia abstract. As our aim is to mine relations for the canonical DBpedia, extracted from the (largest) English language Wikipedia, we inspect all links in the abstract which have a corresponding entity in the main DBpedia knowledge base created from the English Wikipedia.[1] For other languages, we

[1] For this work, we use the 2014 version of DBpedia, which was the most recent release available at the time the experiments were conducted. This version is available at

take one intermediate step via the interlanguage links in Wikipedia, which are extracted as a part of DBpedia [16].

3.1 Overall Approach

For 395 relations that can hold between entities, the ontology underlying the DBpedia knowledge graph[2] defines an explicit domain and range, i.e., the types of objects that are allowed in the subject and object position of this relation.[3] Each Wikipedia page also maps to an entity in the DBpedia knowledge graph, some of which are typed. We consider a pair of a Wikipedia page p_0 and a Wikipedia page p_1 linked from the abstract of p_0 as a *candidate* for a relation R if the corresponding DBpedia entities e_0 and e_1 have types that are equal to the domain and range of R. In that case, $R(e_0, e_1)$ is considered a candidate axiom to be included in the DBpedia knowledge graph. In the example in Fig. 1, given that the *genre* relation holds between musical artists and genres, and the involved entities are of the matching types, one candidate each is generated from both the English and the French DBpedia.[4]

We expect that candidates contain a lot of false positives. For example, for the *birthplace* relation holding between a person and a city, all cities linked from the person's web page would be considered candidates. However, cities may be referred to for various different reasons in an abstract about a person (e.g., they may be their death place, the city of their alma mater, etc.). Thus, we require additional evidence to decide whether a candidate actually represents a valid instantiation of a relation.

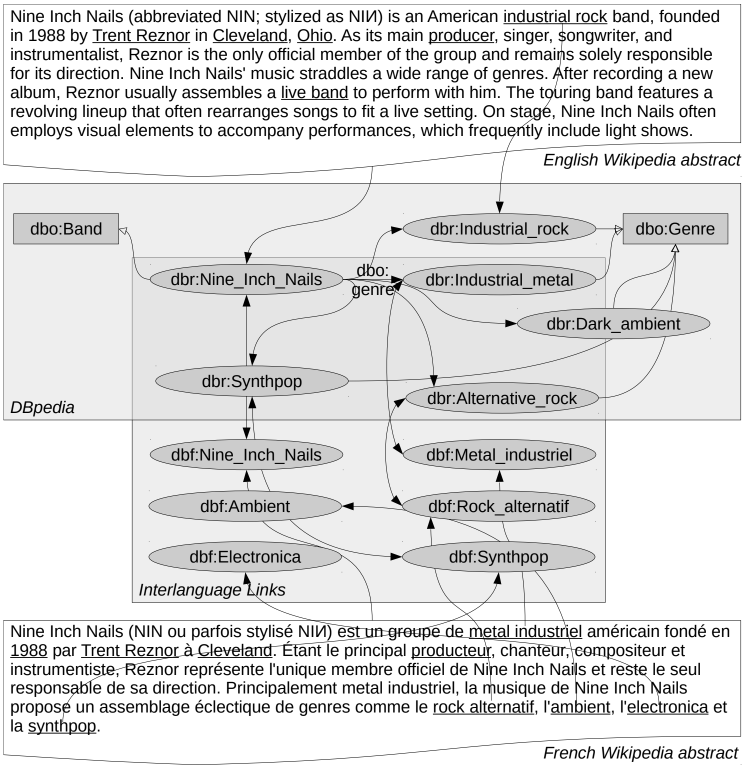
For taking that decision, we train a *machine learning model.* For each abstract of a page for which a given relation is present in the knowledge base, we use the *partial completeness assumption* [11] or *local closed world assumption* [7], i.e., we consider the relation to be complete. Hence, all candidates for the relation created from the abstract which are contained in the knowledge base are considered as *positive* training examples, all those which are not contained are considered as *negative* training examples. In the example in Fig. 1, *Industrial Rock* would be considered a positive example for the relation *genre*, whereas the genre *Rock*, if it were linked in the abstract, would be considered a negative example, since it is not linked as a genre in the DBpedia knowledge graph.

`http://oldwiki.dbpedia.org/Downloads2014`. All statements made in this paper about the size etc. of DBpedia correspond to that version.

[2] `http://dbpedia.org/services-resources/ontology`

[3] Note that the underlying OWL ontology distinguishes *object properties* that hold between entities, and *datatype properties* that hold between an entity and a literal value. Here, we only regard the former case.

[4] Prefixes used in this paper: `dbr=http://dbpedia.org/`, `dbf=http://fr.dbpedia.org/`, `dbo=http://dbpedia.org/ontology/`



Nine Inch Nails (abbreviated NIN; stylized as NIИ) is an American industrial rock band, founded in 1988 by Trent Reznor in Cleveland, Ohio. As its main producer, singer, songwriter, and instrumentalist, Reznor is the only official member of the group and remains solely responsible for its direction. Nine Inch Nails' music straddles a wide range of genres. After recording a new album, Reznor usually assembles a live band to perform with him. The touring band features a revolving lineup that often rearranges songs to fit a live setting. On stage, Nine Inch Nails often employs visual elements to accompany performances, which frequently include light shows.

Nine Inch Nails (NIN ou parfois stylisé NIИ) est un groupe de metal industriel américain fondé en 1988 par Trent Reznor à Cleveland. Étant le principal producteur, chanteur, compositeur et instrumentiste, Reznor représente l'unique membre officiel de Nine Inch Nails et reste le seul responsable de sa direction. Principalement metal industriel, la musique de Nine Inch Nails propose un assemblage éclectique de genres comme le rock alternatif, l'ambient, l'electronica et la synthpop.

Fig. 1: Approach illustrated with extraction from English (above) and French (below) Wikipedia abstract

3.2 Feature Engineering

For training a classifier, both positive and negative examples need to be described by *features*. Table 2 sums up the features used by the classifiers proposed in this paper.

We use features related to the actual candidates found in the abstract (i.e., entities whose type matches the range of the relation at hand), i.e., the total number of candidates in the abstract (F00) and the candidate's sentence (F01), the position of the candidate w.r.t. all other candidates in the abstract (F02) and the candidate's sentence (F03), as well as the position of the candidate's sentence in the abstract (F07). The same is done for all entities, be it candidates or not (F04,F05,F06). Since all of those measures yield positive integers, they are normalized to $(0, 1]$ by using their inverse.

Further features taken into account are the existence of a back link from the candidate's page to the abstract's page (F08), and the vector of all the candi-

Table 1: Example Feature Representation

Instance	F00	F01	F02	F03	F04	F05	F06	F07	F08	`dbo:MusicGenre`	`dbo:Place`	`dbo:Band`	...	Correct
Industrial_Metal	0.2	1.0	1.0	1.0	0.25	1.0	1.0	1.0	1.0	true	false	false	...	true
Alternative_Rock	0.2	0.2	0.5	1.0	0.25	0.2	1.0	0.3	1.0	true	false	false	...	true
Ambient_music	0.2	0.2	0.3	0.5	0.25	0.1	0.5	0.3	0.0	true	false	false	...	false
Electronica	0.2	0.2	0.25	0.25	0.3	0.1	0.3	0.3	0.0	true	false	false	...	false
Synthpop	0.2	0.2	0.2	0.25	0.25	0.1	0.25	0.3	0.0	true	false	false	...	true

Table 2: List of features used by the classifier

ID	Name	Range	ID	Name	Range
F00	NumberOfCandidates	(0,1]	F05	EntityPosition	(0,1]
F01	CandidatesInSentence	(0,1]	F06	EntityPositionInSentence	(0,1]
F02	CandidatePosition	(0,1]	F07	SentencePosition	(0,1]
F03	CandidatePositionInSentence	(0,1]	F08	BackLink	T / F
F04	EntitiesInSentence	(0,1]	FXX	InstanceTypes	T / F

date's types in the DBpedia ontology (FXX).[5] Table 1 depicts the translated feature table for the French Wikipedia abstract depicted in Fig. 1. In this example, there are five candidates (i.e., entities of type `dbo:MusicGenre`), three of which are also contained in the DBpedia knowledge graph (i.e., they serve as true positives).

For the creation of those features which are dependent on the types, the types are taken from the canonical (i.e., English) DBpedia, using the interlanguage links between the language specific chapters, as indicated in Fig. 1.

With the help of a feature vector representation, it is possible to learn fine-grained classification models, such as *The first three genres mentioned in the first or second sentence of a band abstract are genres of that band.*

3.3 Machine Learning Algorithms

Initially, we experimented with a set of five classification algorithms, i.e., Naive Bayes, RIPPER [5], Random Forest (RF) [4], Neural Networks [14] and Support Vector Machines (SVM) [6]. For all those classifiers, we used the implementation in *RapidMiner*[6], and, for the preliminary evaluation, all classifiers were used in their standard setup.

For those five classifiers, we used samples of size 50,000 from the ten most frequent relations in DBpedia, the corresponding English language abstracts, and performed an experiment in ten-fold cross validation. The results are depicted in Table 3. We can observe that the best results in terms of F-measure are achieved by Random Forests, which has been selected as the classifier to use in the subsequent experiments.

Furthermore, we compared the machine learning approach to four simple baselines using the same setup:

[5] The subject's types are not utilized. For DBpedia, they only exist if the subject has an infobox, which would make the approach infeasible to use for long tail entities for which the Wikipedia page does not come with an infobox.

[6] `http://www.rapidminer.com/`

Table 3: Pre-Study results on five machine learning algorithms

	Naive Bayes			Rand.For.			RIPPER			Neural Net			SVM		
Relation	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
`dbo:birthPlace`	.69	.65	.67	.69	.76	.72	.72	.73	.72	.61	.75	.67	.72	.74	.73
`dbo:family`	.55	.93	.69	.87	.83	.85	.85	.83	.84	.77	.83	.80	.87	.83	.85
`dbo:deathPlace`	.42	.30	.35	.51	.30	.38	.64	.18	.28	.61	.19	.29	.66	.20	.31
`dbo:producer`	.35	.55	.43	.35	.14	.20	.47	.04	.07	.23	.10	.14	.48	.05	.09
`dbo:writer`	.55	.61	.58	.62	.55	.58	.64	.54	.59	.52	.51	.51	.67	.53	.59
`dbo:subsequentWork`	.11	.21	.14	.35	.10	.16	.42	.02	.04	.21	.07	.11	.61	.06	.11
`dbo:previousWork`	.18	.43	.25	.39	.18	.25	.59	.05	.09	.57	.08	.14	.60	.10	.17
`dbo:artist`	.94	.94	.94	.94	.95	.94	.95	.96	.95	.95	.86	.90	.95	.89	.92
`dbo:nationality`	.76	.90	.82	.76	.92	.83	.77	.91	.83	.72	.81	.76	.77	.92	.84
`dbo:formerTeam`	.79	.74	.76	.85	.88	.86	.85	.88	.86	.82	.77	.79	.85	.89	.87
Average	.53	.63	.56	.63	.56	.58	.69	.51	.53	.60	.50	.51	.72	.52	.55

Table 4: Pre-Study results on the four baselines

	Baseline 1			Baseline 2			Baseline 3			Baseline 4		
Relation	P	R	F	P	R	F	P	R	F	P	R	F
`dbo:birthPlace`	.47	.99	.64	.46	1.00	.63	.49	.98	.66	.48	.99	.65
`dbo:family`	.18	.85	.30	.17	1.00	.29	.87	.84	.86	.86	1.00	.92
`dbo:deathPlace`	.28	.97	.43	.27	1.00	.43	.30	.93	.46	.30	.96	.46
`dbo:producer`	.17	.93	.29	.15	1.00	.26	.33	.80	.47	.32	.87	.46
`dbo:writer`	.41	.69	.52	.19	1.00	.32	.56	.59	.58	.45	.86	.59
`dbo:subsequentWork`	.02	1.00	.04	.02	1.00	.04	.02	.19	.03	.02	.19	.03
`dbo:previousWork`	.04	1.00	.08	.04	1.00	.08	.04	.20	.06	.03	.20	.06
`dbo:artist`	.31	.99	.47	.27	1.00	.42	.57	.87	.69	.53	.87	.66
`dbo:nationality`	.73	.96	.83	.64	1.00	.78	.74	.96	.84	.64	1.00	.78
`dbo:formerTeam`	.35	.72	.47	.40	1.00	.57	.78	.70	.74	.81	.98	.89
Average	.30	.91	.41	.26	1.00	.38	.47	.71	.54	.44	.79	.55

Baseline 1 The first entity with a matching type is classified as a positive relation, all others as negative.

Baseline 2 All entities with a matching type are classified as positive relations.

Baseline 3 The first entity with a matching ingoing edge is classified as a positive relation. For example, when trying to extract relations for `dbo:birthPlace`, the first entity which already has one ingoing edge of type `dbo:birthPlace` would be classified as positive.

Baseline 4 All entities with a matching ingoing edge are classified as positive relations.

The results of the baseline evaluations are depicted in Table 4. We can observe that in terms of F-measure, they are outperformed by RandomForest. Although the margin seems small, the baseline approaches usually have a high recall, but low precision. In fact, none of them reaches a precision above 0.5, which means that by applying such approaches, at least half of the relations inserted into a knowledge graph would be noise.

4 Experiments

We conducted different experiments to validate the approach. First, we analyzed the performance of the relation extraction using a RandomForest classifier on the English DBpedia only. Here, we follow a two-fold approach: for once, we use a cross-validated silver standard evaluation, where we evaluate how well existing relations can be predicted for instances already present in DBpedia. Since such a silver-standard evaluation can introduce certain biases [22], we additionally validate the findings on a subset of the extracted relations in a manual retrospective evaluation.

In a second set of experiments, we analyze the extraction of relations on the twelve largest language editions of Wikipedia, which at the same time are those with more than 1M articles, i.e., English, German, Spanish, French, Italian, Dutch, Polish, Russian, Cebuano, Swedish, Vietnamese, and Waray.[7,8] Note that this selection of languages does not only contain Indo-European, but also two Austroasiatic and an Austronesian language.

In addition, we conduct further analyses. First, we investigate differences of the relations extracted for different languages with respect to topic and locality. For the latter, the hypothesis is that information extracted, e.g., for places from German abstracts is about places in German speaking countries.

4.1 Pre-Study on English Abstracts

In a first set of experiments, we analyzed the performance of our method on English abstracts only. Since we aim at augmenting the DBpedia knowledge graph at a reasonable level of precision, our aim was to learn models which reach a precision of at least 95%, i.e., that add statements with no more than 5% noise to the knowledge graph. Out of the 395 relations under inspection, the RandomForest classifier could learn models with a precision of 95% or higher for 99 relations. For the 99 models that RF could extract with a minimum precision of 95%, the macro (micro) average recall and precision are 31.5% (30.6%) and 98.2% (95.7%), respectively.

By applying the 99 models to all candidates, a total of 998,993 new relation instances could be extracted, which corresponds to roughly 5% of all candidates. Fig. 2 depicts the 20 relations for which most instances are extracted.

For validating the precision and recall scores computed on the existing relation instances, we sampled each 200 *newly* generated from five relations (i.e., 1,000 in total) and validated them manually. For the selection of entities, we aimed at a wider coverage of common topics (geographic entities, people, books, music works), as well as relations which can be validated fairly well without the need of any specific domain knowledge. The results are depicted in Table 5. It can be observed that the precision values obtained in cross-validation are rather reliable (i.e., the deviation from the estimate is 3% on average), while the recall values are less reliable (with a deviation of 9% on average). The first observation is crucial, as it allows to create new relations for the knowledge graph at

[7] According to `http://wikistats.wmflabs.org/display.php?t=wp`, as of December 2015.

[8] The datasets of the extracted relations for all languages can be found online at `http://dws.informatik.uni-mannheim.de/en/research/language-agnostic-relation-extraction-from-wikipedia-abstracts`

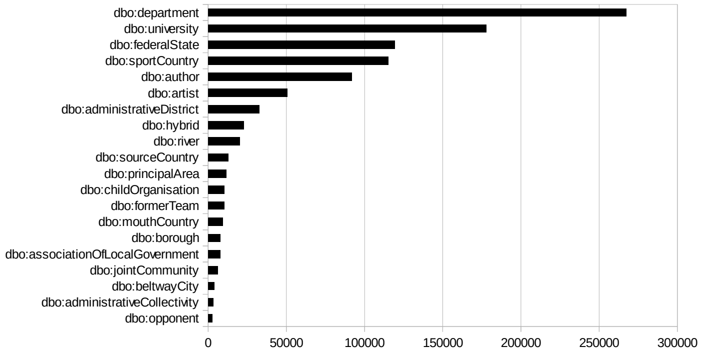


Fig. 2: 20 most frequent relations extracted from English abstracts

Table 5: Results of the manual verification of precision and recall scores computed on the existing relation instances. R_e and P_e denotes the recall and precision of the models computed on the existing relation instances, while R_m and P_m denotes those verified by manual computation.

Relation	R_e	P_e	R_m	P_m
dbo:musicalBand	96.2	95.1	87.9	96.7
dbo:author	68.2	95.2	53.4	91.9
dbo:department	64.5	99.5	53.5	93.7
dbo:sourceCountry	98.9	98.0	98.8	97.8
dbo:saint	41.2	100	53.25	95.5

a reasonable level of precision, i.e., the amount of noise introduced is strictly controlled.

4.2 Cross-Lingual Relation Extraction

In the next experiment, we used the RandomForests classifier to extract models for relations for the top 12 languages, as depicted in Table 6. One model is trained per relation and language.

As a first result, we look at the number of relations for which models can be extracted at 95% precision. While it is possible to learn extraction models for 99 relations at that level of precision for English, that number almost doubles to 187 when using the top twelve languages, as depicted in Fig. 3. These results show that it is possible to learn high precision models for relations in other languages for which this is not possible in English.

When extracting new statements (i.e., instantiations of the relations) using those models, our goal is to extract those statements in the canonical DBpedia knowledge base, as depicted in Fig. 1. The number of extracted statements per language, as well as cumulated statements, is depicted in Fig. 3.

At first glance, it is obvious that, although a decent number of models can be learned for most languages, the number of statements extracted are on average an order of magnitude smaller than the number of statements that are

Table 6: Size of the 12 largest language editions of Wikipedia, and percentage of articles linked to English.

Language	# Entities	% links to English	Language	# Entities	% links to English
English	4,192,414	100.00	Russian	1,277,074	42.61
Swedish	2,351,544	17.60	Waray	1,259,540	12.77
German	1,889,351	42.21	Italian	1,243,586	55.69
Dutch	1,848,249	32.98	Spanish	1,181,096	54.72
French	1,708,934	51.48	Polish	1,149,530	53.70
Cebuano	1,662,301	5.67	Vietnamese	1,141,845	28.68

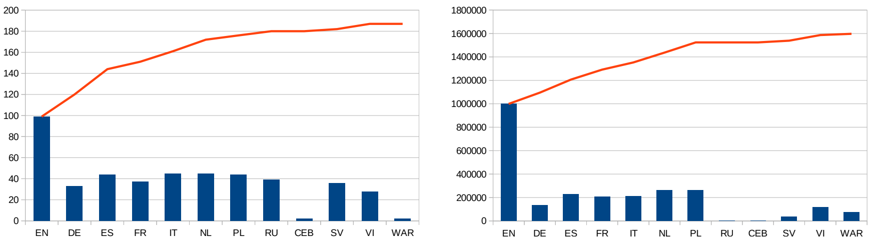


Fig. 3: Number of relations (left) and statements (right) extracted at 95% precision in the top 12 languages. The bars show the number of statements that could be extracted for the given language, the line depicts the accumulated number of statements for the top N languages.

extracted for English. However, the additional number of extracted relations is considerable: while for English only, there is roughly 1M relations, 1.6M relations can be extracted from the top 12 languages, which is an increase of about 60% when stepping from an English-only to a multi-lingual extraction. The graphs in Figure 3 also shows that the results stabilize after using the seven largest language editions, i.e., we do not expect any significant benefits from adding more languages with smaller Wikipedias to the setup.

As can be observed in Fig. 3, the number of extracted statements is particularly low for Russian and Cebuano. For the latter, the figure shows that only a small number of high quality models can be learned, mostly due to the low number of inter-language links to English, as depicted in Table 6. For the former, the number of high quality models that can be learned is larger, but the models are mostly unproductive, since they are learned for rather exotic relations. In particular, for the top 5 relations in Fig. 2, no model is learned for Russian.

It is evident that the number of extracted statements is not proportional to the relative size of the respective Wikipedia, as depicted in Table 6. For example, although the Swedish Wikipedia is more than half the size of the English one, the number of extracted statements from Swedish is by a factor of 28 lower than those extracted from English. At first glance, this may be counter intuitive.

The reason for the number of statements extracted from languages other than English is that we only generate candidates if both the article at hand and the entity linked from that article's abstract have a counterpart in the canonical English DBpedia. However, as can be seen from Table 6, those links to counterparts are rather scarce. For the example of Swedish, the probability of an entity being linked to the English Wikipedia is only 0.176. Thus, the probability for a candidate that both the subject and object are linked to the English Wikipedia is $0.176 \times 0.176 = 0.031$. This is pretty exactly the ratio of statements extracted from Swedish to statements extracted from English (0.036). In fact, the number of extracted statements per language and the squared number of links between the respective language edition and the English Wikipedia have a Pearson correlation coefficient of 0.95. This shows that the low number of statements is mainly an effect of missing inter-language links in Wikipedia, rather than a shortcoming of the approach as such.[9]

4.3 Topical and Geographical Analysis by Language

To further analyze the extracted statements, we look at the topical and geographical coverage for the *additional* statements (i.e., statements that are not yet contained in DBpedia) that are extracted for the twelve languages at hand. First, we depict the most frequent relations and subject classes for the statements. The results are depicted in Fig. 4 and 5. It can be observed that the majority of statements is related to geographical entities and their relations. The Russian set is an exception, since most extracted relations are about musical works, in contrast to geographic entities, as for the other languages. Furthermore, the English set has the largest fraction of person related facts.

We assume that the coverage of Wikipedia in different languages is, to a certain extent, biased towards places, persons, etc. from countries in which the respective language is spoken.[10] Thus, we expect that, e.g., for relations extracted about places, we will observe that the distribution of countries to which entities are related differs for the various language editions.

To validate this hypothesis, we determine the country to which a statement is related as follows: given a statement s in the form

```
s p o .
```

we determine the set of pairs $P_s :=< r, c >$ of relations and countries that fulfill

```
s r c .
c a dbo:Country .
```

and

```
o r c .
c a dbo:Country .
```

For all statements S extracted from a language, we sum up the relative number of relations of a country to each statement, i.e., we determine the weight of a country C as

[9] If we were interested in extending the coverage of DBpedia not only w.r.t. relations between existing entities, but also adding *new* entities (in particular: entities which only exist in language editions of Wikipedia other than English), then the number of statements would be larger. However, this was not in the focus of this work.

[10] See, e.g., `http://geography.oii.ox.ac.uk/?page=geographic-intersections-of-languages-in-wikipedia` for evidence.

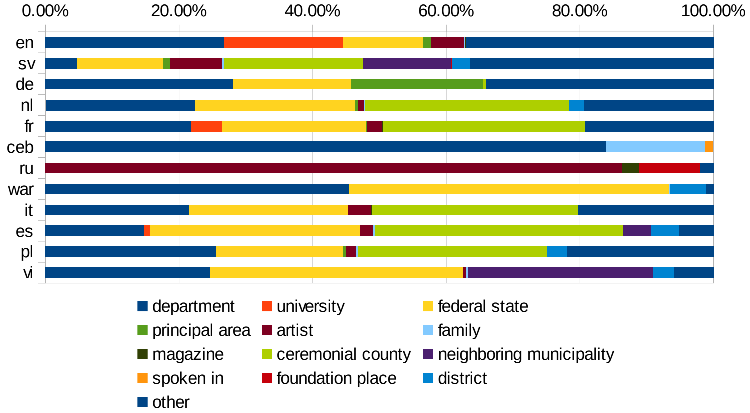


Fig. 4: Distribution of relations in the different language extractions

$$w(C) := \sum_{s=1}^{|S|} \frac{|\{< r, c >\in P_s | c = C\}|}{|P_s|} \tag{1}$$

The analysis was conducted using the RapidMiner Linked Open Data Extension [25].

Figure 6 depicts the distributions for the countries. We can observe that while in most cases, facts about US related entities are the majority, only for Polish, entities related to Poland are the most frequent. For Swedish, German, French, Cebuano and Italian, the countries with the largest population speaking those languages (i.e., Sweden, Germany, France, Philippines, and Italy, respectively), are at the second position. For Spanish, Spain is at the second position, despite Mexico and Colombia (rank 11 and 6, respectively) having a larger population. For the other languages, a language-specific effect is not observable: for Dutch, the Netherlands are at rank 8, for Vietnamese, Vietnam is at rank 34, for Waray, the Philippines are at rank 7. For Russian, Russia is on rank 23, preceded by Soviet Union (sic!, rank 15) and Belarus (rank 22).

The results show that despite the dominance of US-related entities, there is a fairly large variety in the geographical coverage of the information extracted. This supports the finding that adding information extracted from multiple Wikipedia language editions helps broadening the coverage of entities.

5 Conclusion and Outlook

Adding new relations to existing knowledge graphs is an important task in adding value to those knowledge graphs. In this paper, we have introduced an approach that adds relations to DBpedia using abstracts in Wikipedia. Unlike other works in that area, the approach presented in this paper uses background knowledge from DBpedia, but does not rely on any language-specific techniques, such as POS tagging, stemming, or dependency parsing. Thus, it can be applied to Wikipedia abstracts in any language.

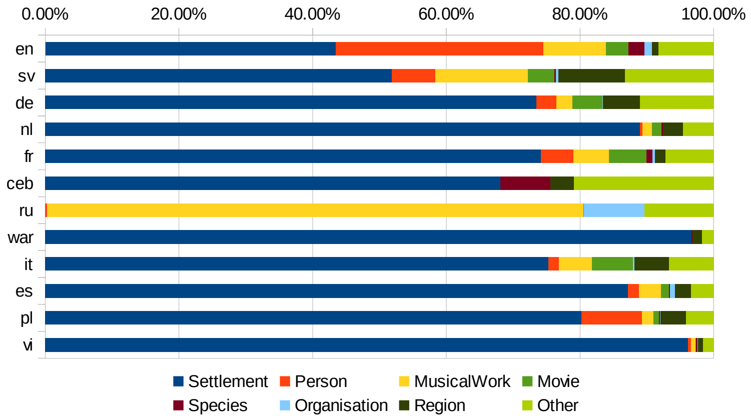


Fig. 5: Distribution of subject types in the different language extractions

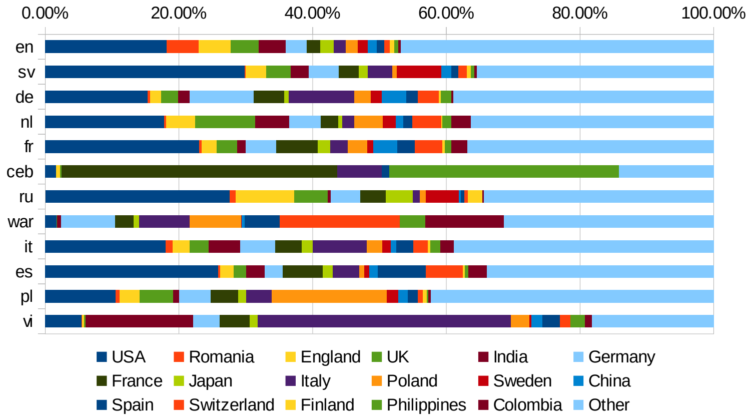


Fig. 6: Distribution of locality in the different language extractions

While we have worked with DBpedia only in this paper, the approach can be applied to other cross-domain knowledge graphs, such as YAGO or Wikidata, as well, since they also link to DBpedia. Furthermore, for a significant portion of Semantic Web datasets, links to DBpedia exist as well [26], so that the approach can be applied even to such domain-specific datasets.

The experimental results show that the approach can add a significant amount of new relations to DBpedia. By extending the set of abstracts from English to the most common languages, the coverage both of relations for which high quality models can be learned, as well as of instantiation of those relations, significantly increases.

Following the observation in [29] that multi-lingual training can improve the performance for each single language, it might be interesting to apply models also on languages on which they had not been learned. Assuming that certain patterns exist in many languages (e.g., the first place being mentioned in an article about a person being the person's birth place), this may increase the amount of data extracted.

In our experiments, we have only concentrated on relations between entities so far. However, a significant fraction of statements in DBpedia and other knowledge graphs also have literals as objects. That said, it should be possible to extend the framework to such statements as well. Although numbers, years, and dates are usually not linked to other entities, they are quite easy to detect using, e.g., regular expressions or specific taggers such as *HeidelTime* [28]. With such a detection step in place, it would also be possible to learn rules for datatype properties, such as: *the first date in an abstract about a person is that person's birthdate*, etc.

Furthermore, our focus so far has been on adding missing relations. A different, yet related problem is the detection of wrong relations [22–24]. Here, we could use our approach to gather *evidence* for relations in different language editions of Wikipedia. Relations for which there is little evidence could then be discarded (similar to DeFacto [12]). While for adding knowledge, we have tuned our models towards *precision*, such an approach, however, would require a tuning towards *recall*. In addition, since there are also quite a few errors in numerical literals in DBpedia [9, 32], an extension such as the one described above could also help detecting such issues.

So far, we have worked on one genre of text, i.e., abstracts of encyclopedic articles. However, we are confident that this approach can be applied to other genres of articles as well, as long as those follow typical structures. Examples include, but are not limited to: extracting relations from movie, music, and book reviews, from short biographies, or from product descriptions. All those are texts that are not strictly structured, but expose certain patterns. While for the Wikipedia abstracts covered in this paper, links to the DBpedia knowledge graph are implicitly given, other text corpora would require entity linking using tools such as DBpedia Spotlight [18].

In summary, we have shown that Wikipedia abstracts are a valuable source of knowledge for extending knowledge graphs such as DBpedia. Those abstracts expose patterns which can be captured by language-independent features, thus allowing for the design of language-agnostic systems for relation extraction from such abstracts.

References

1. Aprosio, A.P., Giuliano, C., Lavelli, A.: Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In: NLP&DBpedia. CEUR Workshop Proceedings, vol. 1064 (2013)
2. Bender, E.M.: Linguistically naïve!= language independent: why NLP needs linguistic typology. In: EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? pp. 26–32 (2009)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In: 2008

ACM SIGMOD International Conference on Management of Data. pp. 1247–1250. ACM (2008)

4. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001), `http://dx.doi.org/10.1023/A:1010933404324`
5. Cohen, W.W.: Fast Effective Rule Induction. In: Machine Learning, Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995. pp. 115–123 (1995)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2010)
7. Dong, X.L., Murphy, K., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Strohmann, T., Sun, S., Zhang, W.: Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In: 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610 (2014)
8. Faruqui, M., Kumar, S.: Multilingual open relation extraction using cross-lingual projection. arXiv preprint arXiv:1503.06450 (2015)
9. Fleischhacker, D., Paulheim, H., Bryl, V., Völker, J., Bizer, C.: Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection. In: The Semantic Web–ISWC 2014, LNCS, vol. 8796, pp. 357–372. Springer (2014)
10. Fundel, K., Küffner, R., Zimmer, R.: RelExRelation extraction using dependency parse trees. Bioinformatics 23(3), 365–371 (2007)
11. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In: 22nd international conference on World Wide Web. pp. 413–422 (2013)
12. Gerber, D., Esteves, D., Lehmann, J., Bühmann, L., Usbeck, R., Ngomo, A.C.N., Speck, R.: DeFacto – Temporal and multilingual Deep Fact Validation. Web Semantics: Science, Services and Agents on the World Wide Web 35(2) (2015)
13. Gerber, D., Ngomo, A.C.N.: Bootstrapping the Linked Data web. In: Workshop on Web Scale Knowledge Extraction (2011)
14. Kubat, M.: Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. Knowledge Eng. Review 13(4), 409–412 (1999), `http://journals.cambridge.org/action/displayAbstract?aid=71037`
15. Lange, D., Böhm, C., Naumann, F.: Extracting structured information from Wikipedia articles to populate infoboxes. In: 19th ACM Conference on Information and Knowledge Management (CIKM). pp. 1661–1664. ACM (2010)
16. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal 6(2) (2013)
17. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A Knowledge Base from Multilingual Wikipedias. In: Conference on Innovative Data Systems Research (2015)
18. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: 7th International Conference on Semantic Systems (2011)
19. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011. Association for Computational Linguistics (2009)
20. Nguyen, D.P., Matsuo, Y., Ishizuka, M.: Relation extraction from wikipedia using subtree mining. In: National Conference on Artificial Intelligence. vol. 22, p. 1414 (2007)

21. Nguyen, T.H., Grishman, R.: Relation extraction: Perspective from convolutional neural networks. In: Proceedings of NAACL-HLT. pp. 39–48 (2015)
22. Paulheim, H.: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web (2016), to appear
23. Paulheim, H., Bizer, C.: Improving the Quality of Linked Data Using Statistical Distributions. International Journal on Semantic Web and Information Systems (IJSWIS) 10(2), 63–86 (2014)
24. Paulheim, H., Gangemi, A.: Serving DBpedia with DOLCE–More than Just Adding a Cherry on Top. In: International Semantic Web Conference. LNCS, vol. 9366. Springer (2015)
25. Ristoski, P., Bizer, C., Paulheim, H.: Mining the web of linked data with rapidminer. Web Semantics: Science, Services and Agents on the World Wide Web 35, 142–151 (2015)
26. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data Best Practices in Different Topical Domains. In: International Semantic Web Conference (2014)
27. Schutz, A., Buitelaar, P.: Relext: A tool for relation extraction from text in ontology extension. In: The Semantic Web–ISWC 2005, pp. 593–606. Springer (2005)
28. Strötgen, J., Gertz, M.: HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In: 5th International Workshop on Semantic Evaluation. pp. 321–324 (2010)
29. Verga, P., Belanger, D., Strubell, E., Roth, B., McCallum, A.: Multilingual relation extraction using compositional universal schema. arXiv preprint arXiv:1511.06396 (2015)
30. Vrandečić, D., Krötzsch, M.: Wikidata: a Free Collaborative Knowledge Base. Communications of the ACM 57(10), 78–85 (2014)
31. Wang, G., Yu, Y., Zhu, H.: PORE: Positive-only Relation Extraction from Wikipedia Text. In: 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference. pp. 580–594. ISWC'07/ASWC'07, Springer-Verlag (2007)
32. Wienand, D., Paulheim, H.: Detecting Incorrect Numerical Data in DBpedia. In: The Semantic Web: Trends and Challenges, LNCS, vol. 8465, pp. 504–518. Springer (2014)
33. Wu, F., Hoffmann, R., Weld, D.S.: Information extraction from Wikipedia: Moving down the long tail. In: 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 731–739. ACM (2008)
34. Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., Ishizuka, M.: Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. In: Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. pp. 1021–1029. ACL '09, Association for Computational Linguistics (2009)
35. Zelenko, D., Aone, C., Richardella, A.: Kernel methods for relation extraction. The Journal of Machine Learning Research 3, 1083–1106 (2003)
36. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal. pp. 17–21 (2015)
37. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., et al.: Relation Classification via Convolutional Deep Neural Network. In: COLING. pp. 2335–2344 (2014)