Multilingual Expert Search using Linked Open Data as Interlingual Representation

Daniel M. Herzig and Hristina Taneva

Institute AIFB
Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
herzig@kit.edu, hristina.taneva@student.kit.edu

Abstract. Most Information Retrieval models take documents as *Bag-of-Words* and are thereby bound to the language of the documents. In this paper, we present an approach using Linked Open Data resources, i.e. URIs, as interlingual document representations. Documents and queries are summarized by the resources they contain. We show the applicability of our approach for multilingual retrieval with a case study on expert search.

1 Introduction

When encountering a problem, there are often two ways to get to a solution. Either acquire the knowledge, in order to solve the problem by oneself, or ask for outside help, preferably somebody who has expertise and experience in the needed domain. The first case is often not feasible or would require too much time. In the second case, the subsequent problem of finding the right expert arises. We address this problem and present an approach for expert search in this paper.

Identifying who an expert is for a certain domain can be done in many ways. One possible solution is to use documents and assume that the authors have expertise on the topics they wrote about. We apply this assumption and consider documents for the identification of experts.

Obviously, the more specific the problem is the harder is it to find an expert. Thus, extending the considered search space even across languages improves the situation. The scenario of considering documents in different languages is not an artificial one, e.g. global companies have product documentations in many languages or online developer forums have discussion threads in different languages. Our approach addresses the problem of how to deal with different languages by applying an interlingual representation for documents based on Linked Open Data resources.

Expert Search Track at CriES Our approach participated at the expert search track of the Cross-lingual Expert Search Workshop (CriES) at CLEF 2010 [18]. The setting and the evaluations presented in this paper are provided by the workshop. The task of the expert search track was to find experts for 60 topics

consisting of 15 topics in each of the four languages English, Spanish, French, and German in the Yahoo! Answers data corpus [19][1]. The data corpus consists of 780193 threads, i.e. questions and answers, in the categories "Health", "Computer & Internet", and "Science & Math.", in four languages written by 169819 users, i.e. experts. Table 1 gives an overview of the data set. More details and an overview of the results of the workshop can be found in [18].

	Threads	Users
English	712370 (91%)	149410 (88%)
Spanish	38722 (5%)	11931 (7%)
French	19867 (3%)	5749 (3%)
German	9234 (1%)	3152 (2%)

Table 1. Overview of the data corpus regarding the size and language distribution.

This paper is organized as follows. After the introduction in this section, we describe the usage of Linked Open Data as an interlingual representation in Section 2. In Section 3, we present our model for expert search, how we create profiles between resources and experts and how we estimate parameters. Section 4 presents the evaluation and Section 5 discusses related work. Finally, we conclude in Section 6.

2 Multilingual IR based using Linked Open Data

Most common models in Information Retrieval see documents as *Bag-of-Words*, i.e. they resolve the order of the words and take the collection of words as the representation of a document. As a consequence, this representation is directly bound to the language of the document. When using keyword queries in one language, relevant documents in another languages are probably not retrieved. We propose an approach using Linked Open Data resources as document representation. Linked Open Data (LOD) refers to interlinked, publicly available, and structured datasets on the web using semantic web standards, in particular the Resource Description Framework (RDF) [4, 7].

The first principle of LOD states that *things* should be identified by Uniform Resource Identifiers (URIs), where *things*, i.e. resources, can be virtually everything. The notion is not limited to physical things, but comprises also abstract or intangible concepts, like *happiness* or *fire alarm*. URIs are not necessarily human readable, since they are meant to be processed by machines. Therefore, human readable labels are often assigned to URIs. Since there can be multiple labels in different languages for one URI, the URI itself can be seen as an interlingual representation for the resource it identifies. Figure 1 illustrates an example about the resource representing *Germany* and its labels in several languages.

The resource in Figure 1 is taken from DBpedia[2]. DBpedia is a popular LOD dataset extracted from Wikipedia, which exploits the interlanguage links

[1] This dataset is provided by the Yahoo! Research Webscope program (see http://research.yahoo.com/) under the following ID:*L6. Yahoo! Answers Comprehensive Questions and Answers (version 1.0)*

[2] http://dbpedia.org, Aug 4 2010

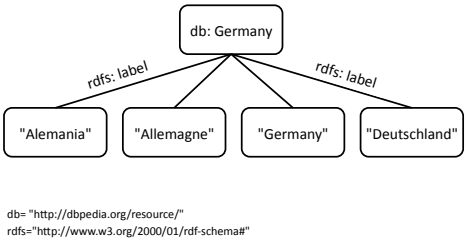

Fig. 1. The resource representing "Germany" with human readable labels in different languages.

of Wikipedia for the labels. Since not all articles have a corresponding article in all other languages, some resources do not have labels in all languages. Figure 2 gives an overview of the number of articles in the considered languages, which directly corresponds to the number of resources and their labels. We use these Wikipedia resources in our approach to capture the *aboutness* [10] of documents. However, our approach is not limited to resources from Wikipedia. Other LOD and RDF resources could be used likewise, e.g. AGROVOC[3], a conceptualization of the agricultural domain features labels in five languages and could be used for documents in this domain.

We used the Wikipedia Miner Toolkit to extract the resources from documents[13]. The miner identifies possible candidates in the text and then disambiguates and verifies them up to a given confidence value by using the link structure of Wikipedia and the surrounding terms, see [12] for details. The extracted resources form a *Bag-of-Resources* representation of the document as illustrated in Figure 3. Each resource identifies unambiguously one thing. As mentioned above, these resources are interlingual even though they have often English names.

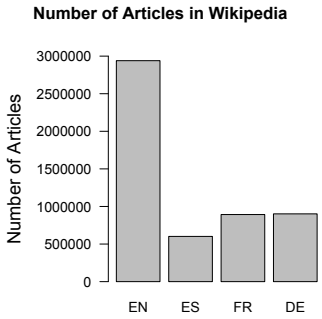

Fig. 2. Number of articles in Wikipedia for different languages as of September 2009.

Although the advantage of Linked Open Data is the connection between resources, we omit this feature and leave the exploitation of links between resources for future work. For now, we use only the resources. Therefore, the current approach seems similar to concept based IR approaches, especially since Wikipedia has been frequently used as a concept space[5], but also EuroWordNet[4] or UWN[6]. However, the difference is that using resources allows to directly

[3] http://www.fao.org/agrovoc/, Aug 4 2010
[4] http://www.illc.uva.nl/EuroWordNet/, Aug 4 2010

exploit additional information from other Linked Open Data sources, e.g. the resource representing *Germany* from Figure 1 is linked through the typed link `owl : sameAs` [2] to resources representing the same thing, e.g. to the resource from *The New York Times*[5] or from *Geonames*[6]. Beside information about the same thing, also the typed links between different resources can be exploited. It allows to enrich the representation with additional information and adapt it to specific use cases or domains.

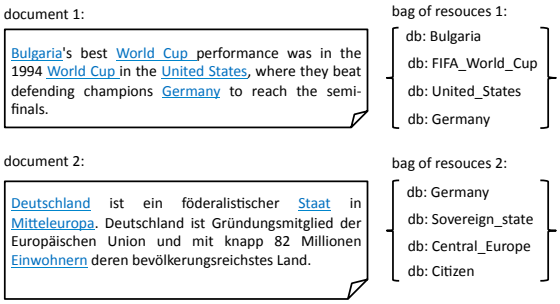

Fig. 3. Text documents in different languages and their interlingual representation in Linked Open Data resources.

3 Expert Search

The Yahoo! Answers data corpus contains discussion threads consisting of an initial question and subsequent answers. The problem of expert search in this context is to find users, who are likely able to answer a given question q, i.e. a topic, based on the threads in the data corpus.

We apply mixture language models. Potential experts are ranked according to the probability that the expert $ex \in E$ can answer the given question $q \in Q$, i.e. $P(ex|q)$. A question q is modeled as a *Bag-of-Resources*: $q = \{r_1, ..., r_n\}$.

$$P(ex|q) \propto P(ex) \cdot P(q|ex) = P(ex) \cdot \prod_{i=1}^{n} P(r_i|ex) \tag{1}$$

We apply Bayes's theorem and assume $P(q)$ and the *prior* $P(ex)$ to be equal to 1. The probability $P(r_i|ex)$ is approximated as a weighted sum of several features f and smoothed by information over the entire corpus C.

$$P(r_i|ex) = \sum_f (\lambda_f \cdot P_f(r_i|ex)) + \lambda_C \cdot P_C(r_i)$$

$$s.t. \sum_f \lambda_f + \lambda_C = 1$$

[5] http://data.nytimes.com/55761554936313344161, Aug 4 2010
[6] http://sws.geonames.org/2921044/about.rdf, Aug 4 2010

3.1 Expert - Resource Profiles

One answer per thread is marked by the questioner or by votes of other users as the best answer to the question. The user, who gave the best answer, is identifiable by its ID. All other answers do not have a user ID. We exploit this setting by building two different models. These models are illustrated in Figure 4 and explained below.

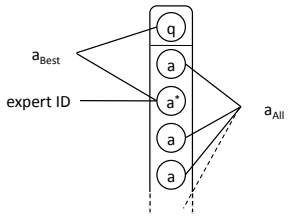

Fig. 4. One example discussion thread. The initial question q and the best answer a^* are combined to a_{best} and subject to the *Best-Answer Model*. All other answers are put together as a_{all} and considered by the *All-other-Answer Model*.

Best-Answer Model
This model takes the question q and the best answer a^* together as a_{best} and relates a_{best} to the expert who gave the best answer, as illustrated in Figuren 4. The idea behind this model is that the user obviously understood the question, because he was able to give the best answer. Therefore he holds expertise about the covered resources. Formally, the model is defined as follows, where $freq(r, a)$ is the frequency of resource r in a.

$$P_{best}(r|ex) = \sum_{a_{best}} P(r|a_{best}) \cdot \frac{P(ex|a_{best}) \cdot P(a_{best})}{P(ex)}$$

with

$$\begin{aligned} P(r|a_{best}) &= \frac{freq(r, a_{best})}{\sum_{r \in a_{best}} freq(r, a_{best})} \\ P(ex|a_{best}) &= 1, \text{iff ex author of } a_{best}, \text{ 0 otherwise} \\ P(a_{best}) &= \frac{1}{|Q|}, P(ex) = \frac{1}{|E|} \end{aligned}$$

All-other-Answers Model
This model relates all answers a_{all}, except the best answer, to the expert, who gave the best answer. The assumption behind this model is that an expert, who gave the best answer, might also say that other answers are not correct. Therefore, we assume that the expert has expertise about the resources covered

by these answers as well, at least to some extent. Formally, the model is defined analogously to the previous one:

$$P_{all}(r|ex) = \sum_{a_{all}} P(r|a_{all}) \cdot \frac{P(ex|a_{all}) \cdot P(a_{all})}{P(ex)}$$

3.2 Parameter Estimation

The mixture model presented in the previous section allows to balance the influence of each model through the corresponding parameter λ_f. In our case, we need to determine λ_{best}, the weight for the Best-Answer-Model and λ_{all}, the weight for the All-other-Answers-Model. The smoothing parameter λ_C remains fixed at $\lambda_C = 0.1$. Hence, $\lambda_{best} + \lambda_{all} \stackrel{!}{=} 0.9$ must hold.
In order to examine the effect of different parameter configurations on the performance of the retrieval, we used the 60 topics provided by workshop along with the given *a priori* relevance information. The *a priori* relevance information is directly taken from the data set, i.e. each topic has exactly one relevant expert, namely the one, who wrote the best answer for this topic. This setting is not optimal, since these questions are part of the data corpus itself and not distinct from it. Furthermore, it can be assumed that there are more than one relevant expert per question and that judging the performance by the occurrence of just one expert in the result set will deviate from the actual result. However, it allows at least to roughly estimate a parameter configuration. We used the *Mean Average Precision (MAP)* to measure the performance.

Since the 60 topics are part of the Best-Answer-Model a correlation between λ_{best} and the MAP can be assumed in this setting. The MAP was measured in steps of 0.05 from $\lambda_{best} = 0$, i.e. the performance without the Best-Answer-Model, to $\lambda_{best} = 0.9$, the performance of the Best-Answer-Model alone. The observed MAP for the different parameters is shown in the left plot of Figure 5. As assumed, a correlation between λ_{best} and the MAP can be observed. Remarkably, the MAP decreases for $\lambda_{best} = 0.9$, despite the assumed correlation. This suggests that information is lost, if the All-Answers-Model is not involved and as a consequence, the optimal parameter configuration can not be the maximal observed MAP. We used least-square curve fitting to approximate the observed values, i.e. the red line in Figure 5. The maximum of the fitted curve is $\lambda_{best} = 0.66$. Comparing the estimated values with the actual MAP computed *ex post* with the entire assessments shows that the maximum is even lower at about $\lambda_{best} = 0.53$, see right plot of Figure 5.

4 Evaluation

We submitted three runs with different configurations, see Table 2 for an overview of the results. Some of the 60 topics are very short and many are written in rather colloquial language and grammar or use abbreviations, e.g. *"Why*

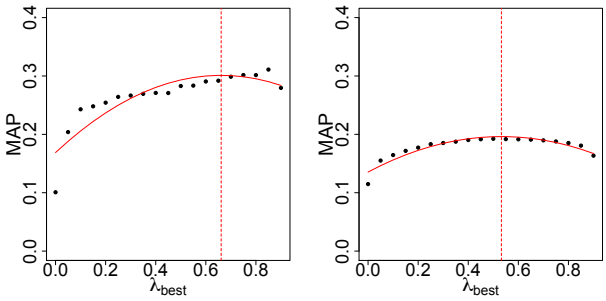

Fig. 5. Parameter estimation through non linear curve fitting (red curve) over the Mean Average Precision for a parameter sweep on λ_{best} (black dots). The left plot shows the *a priori* estimation computed with the relevant information about the best expert only. The right plot shows the actual MAP computed *ex post* with the entire assessments.

do women get PMS?" or *"hab es runtergeladen wie kann ich bei msn chatten?"*, which caused problems for the Wikipedia Miner to identify resources. For 11 topics the Wikipedia Miner did not identify any resources. In these cases, we extracted the resources manually, e.g. the resources *db:Woman* and *db:Premenstrual_syndrome* for the first question mention before and the resources *db:MSN* and *db:Online_chat* for the latter. We did this for run1 and run3 and left the topics untouched for run2, in order to see how the approach performs without any manual intervention.

	Strict		Lenient		Parameters		
Run Id	P@10	MRR	P@10	MRR	λ_{best}	λ_{all}	λ_C
run3	0.49 (+157%)	0.76 (+90%)	0.87 (+123%)	0.93 (+48%)	0.7	0.2	0.1
run1	0.48 (+153%)	0.77 (+93%)	0.86 (+121%)	0.94 (+49%)	0.6	0.3	0.1
run2	0.35 (+84%)	0.65 (+63%)	0.61 (+56%)	0.74 (+17%)	0.6	0.3	0.1
BM25 + Z-Score	0.19	0.40	0.39	0.63			

Table 2. Results of the runs submitted to the CriES pilot challenge. The percentages show the performance against the BM25+Z-Score baseline.

Table 2 shows the results for the top 10 retrieved experts. Precision at cut-off level 10 (P@10) and Mean Reciprocal Rank (MRR) are used as evaluation measures. Precision/Recall curves for each run are presented in Figure 6 using strict and lenient assessments [18]. All three runs exceed the standard IR baseline, BM25 + Z-Score [18]. The baseline uses machine translation to translate the topics in the four languages and matches them against monolingual indexes. The results retrieved from the four monolingual indexes are combined for each expert using the Z-Score [15].

Beside retrieving relevant experts for a topic, one main aim of our approach was to cross the language barrier and find experts regardless of their language. Figure 7 visualizes the language distribution of the retrieved experts for each topic language for run1. In order to facilitate the comparison, the distribution

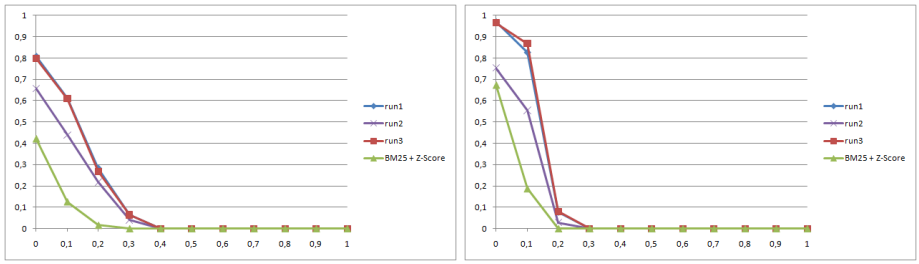

Fig. 6. Precision/Recall curves based on interpolated recall for strict (left plot) and lenient (right plot) assessment.

of threads and experts in the data set is displayed on the right. One can see that indeed experts in all four languages were retrieved in most cases. Further, the domination of english speaking experts is due to the proportions in the data set and in addition due to the larger, underlying resource space, as illustrated in Figure 2. However, a bias towards the language of the topic is also observable, because not all resource have labels in all other languages as discussed in Section 2.

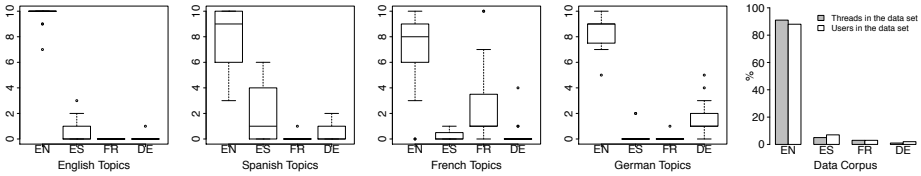

Fig. 7. Boxplots illustrating the distribution of the top 10 retrieved experts by language for each topic language. The right most plot shows the distribution by language of threads and experts in the Yahoo Answers data set.

5 Related Work

Our approach uses Language Models, which have been studied by [9] and applied by [3] for expert search. The latter compares two models with different search strategies. The first model collects all documents for every candidate and then identifies the relevant topics in these documents. The second model finds first the significant documents for a given topic and then discovers the associated experts. Our approach is based on model similar to the second model. We find first the documents comprising the resources of the query and then relate the resources to the expert who gave the best answer.

Using concepts instead of terms was studied by [14, 16, 17]. These approaches use Explicit Semantic Analysis and match topics to documents in a concept space consisting of Wikipedia articles. Our approach uses also Wikipedia as background knowledge and a representation similar to the concept representation, as long as only the resources are considered. However, as discussed in Section 2, using URIs instead of concepts allows to draw information from other sources

and facilitates the usage of connections between the resources. In order to extract the resources from the documents, we screened several tools that deal with Named Entity Recognition and Extraction. The Enrycher web service [21] tries to extract not just resources, but also triples, which connect these resources. The OpenCalais Web Service analyzes text and returns semantic metadata[1]. However, these approaches do not work with all four languages. The advantage of the Wikipedia Miner [13], beside fast and precise results, is that the resource space is clearly defined, i.e. all articles of Wikipedia, and that it supports the language of the loaded Wikipedia file. Hence, we choose [13] for our approach.

Not just indexing the terms of a document, but the idea of indexing what a document is *about*, i.e. *topic indexing*, was introduced by [10]. Another approach to topic indexing by embedding background knowledge derived from Wikipedia was introduced by [11]. All relevant topics being mentioned in a document are linked to Wikipedia articles. The titles of the articles are used as index terms. A similar approach to ours, however our approach is not limited to Wikipedia and the usage of URIs instead of terms allows to exploit the links between the URIs in a later stage.

A different approach to multilingual IR was introduced by [8], who uses a multilingual ontology to map a term with the appropriate concept. However, it does not consider disambiguation of terms. An aspect covered by our approach, since URIs are not ambiguous and the URIs are determined using the disambiguation of [13]. [20] examines the impact of the use of semantic annotations on the performance in monolingual and cross-language retrieval.

6 Conclusions and Future Work

We presented an approach for the Expert Search challenge of the CriES Workshop at CLEF 2010 using Linked Open Data resources, i.e. URIs, as interlingual document representations. We used Wikipedia as the corpus of resources, but the approach is not limit to the usage of Wikipedia. Resources are extracted from the documents using the Wikipedia Miner Toolkit [13] and used to create Expert-Resource profiles. A mixture model is applied for the retrieval and ranking of experts for a given topic. Also topics are represented as a *Bag-of-Resources*.

Our approach yielded solid results by exceeding the standard BM25 + Z-Score baseline from 17% to 157% regarding Mean Reciprocal Rank and Precision at 10. Another advantage of our approach is that not the entire documents need to be indexed, but just a summary consisting of several URIs, which decreases the index size.

In future, we plan to use more features of Linked Open Data for IR. In particular, exploiting the links between resources to leverage the interconnection.

7 Acknowledgments

We thank Philipp Sorg for the helpful discussions and his valuable feedback. We also thank David Milne for developing the Wikipedia Miner and making it available as open source. Research reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF) under the iGreen project (grant 01IA08005K).

References

1. Opencalais. http://www.opencalais.com, Aug 4 2010.
2. Owl web ontology language overview. W3c recommendation, World Wide Web Consortium, February 2004.
3. K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Information Processing & Management*, 45(1):119, 2009.
4. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
5. P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit vs. latent concept models for cross-language information retrieval. In *Proceedings of the Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 1513–1518. AAAI Press, July 2009.
6. G. de Melo and G. Weikum. Towards a universal wordnet by learning from combined evidence. In *Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA, 2009. ACM.
7. E. M. Frank Manola. RDF primer. http://www.w3.org/TR/rdf-primer/, Feb. 2004. W3C Recommendation.
8. J. Guyot, S. Radhouani, and G. Falquet. Ontology-based multilingual information retrieval. In *CLEF Workhop, Working Notes Multilingual Track*, pages 21–23, 2005.
9. D. Hiemstra. Using language models for information retrieval. *University of twente*, 2001.
10. M. Maron. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, pages 38–43, 1977.
11. O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *Proc. of the AAAI WikiAI workshop*, 2008.
12. D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proc. of the 17th ACM Conf. on Information and knowledge management (CIKM)*, pages 509–518, Napa Valley, California, USA, 2008. ACM.
13. D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. In *Proc. New Zealand Computer Science Research Student Conf.*, volume 9, 2009.
14. M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *ECIR*, pages 522–530, 2008.
15. J. Savoy. Data fusion for effective european monolingual information retrieval. In *Multilingual Information Access for Text, Speech and Images*, pages 233—244. 2005.
16. P. Sorg and P. Cimiano. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*, 2008.
17. P. Sorg and P. Cimiano. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Proc. of the Int. Conf. on Applications of Natural Language to Information Systems (NLDB)*, pages 36–48. Springer, June 2009.
18. P. Sorg, P. Cimiano, and S. Sizov. Overview of the cross-lingual expert search (CriES) pilot challenge. In *Working Notes of the CLEF 2010 Lab Sessions*, 2010.
19. M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers on large online QA collections. In *Proc. of the 46th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, page 719727, 2008.
20. M. Volk and P. Buitelaar. Ontologies in cross-language information retrieval. In *Proceedings of WOW2003*, 2003.
21. T. Štajner, D. Rusu, L. Dali, B. Fortuna, D. Mladenić, and M. Grobelnik. EnrycherService oriented Text Enrichment. *Proc. of SiKDD*, 2009.