Renato Domínguez García, Philipp Scholl, Christoph Rensing: Supporting Ressource-based Learning on the Web using automatically extracted Large-scale Taxonomies from multiple Wikipedia versions. In: Howard Leung, Elvira Popescu, Yiwei CaoRynson W.H. Lau, Wolfgang Nejdl: Advances in Web-based Learning, Proceedings of the 10th International Conference on Web-based Learning, ICWL 2011, LNCS 7048, pp. 309-314, Springer, December 2011. ISBN: 978-3642258121.

Supporting Resource-based Learning on the Web using automatically extracted Large-scale Taxonomies from multiple Wikipedia versions

Renato Domínguez García, Philipp Scholl, and Christoph Rensing

Multimedia Communications Lab - Technische Universität Darmstadt
64283 Darmstadt - Germany
{renato,scholl,rensing}@kom.tu-darmstadt.de
http://www.kom.tu-darmstadt.de

Abstract. CROKODIL is a platform for the support of collaborative resource-based learning with web resources. It enables the building of learning communities in which learners annotate their relevant resources using tags. In this paper, we propose the use of automatically generated large-scale taxonomies in different languages to cope with two challenges in CROKODIL: The multilingualism of the resources, i.e. web resources are in different languages and the connectivity of the semantic network, i.e. learners do not tag resources on the same topic with identical tags. More specifically, we describe a set of features that can be used for detecting hyponymy relations from the category graph of Wikipedia. Finally, we evaluate our approach on Wikipedia versions in four different languages, namely English, German, Spanish and Arabic.

Keywords: Resource-based Learning, Hyponymy Detection, Wikipedia Mining, TEL Recommender

1 Introduction

Learning is becoming a lifelong process which has to be continuously performed due to changing working environments and decreased life-span of knowledge. Learners are themselves now responsible for their own learning processes and are becoming free to decide what, when, where and how they want to learn. As the importance of the World Wide Web as a major source for knowledge acquisition has been growing steadily over the last decade, web resources are available on a large scale. Self-directed learning using learning resources is called *Resource–Based Learning*. In Resource-based Learning on the web, one challenge for learners is to find relevant web resources. Web search engines or digital libraries are common tools to support this task. However, in learning settings, where a community like a learning group or a research group exists, there is a high probability that relevant resources were already found.

Recommender systems can be helpful in order to find these resources. However, the domain of Technology Enhanced Learning (TEL) have special requirements. For example, in TEL there are different audiences in different stages of

achieved expertise needing different types of learning materials: Novices need resources giving a broad overview of the learning domain, whereas experts need resources having a very narrow scope. Further, the different levels of knowledge are reflected by the used terminology. For instance, novices tend not to be aware of terminology of the domain they are interested in, while experts are able to communicate in a brief manner using the professional terminology.

A taxonomy of topic may be helpful to recognize general and specific tags and resources and provide support of learners in knowledge acquisition. This goal should be achieved by deriving a taxonomy of topics using a machine learning approach with a set of features from the category system of Wikipedia. The Wikipedia Category Graph provides pairs of related concepts whose semantic relation is unspecified [10]. Our set of features is used to differentiate between *is-a* (taxonomic) and *not-is-a* (all other kind) relations. In section 2, we describe our application scenario. Related work is presented in section 3. In section 4, we describe our proposed approach. The results of our evaluation using different languages are presented in section 5. Finally, we draw conclusions in section 6.

2 Our Application Scenario

2.1 Introducing the CROKODIL platform

The scenario we address with this work is self-directed, collaborative, resource-based learning on the web using a a platform called CROKODIL [11]. It supports learners in acquiring knowledge based on web resources and combines functionalities for the management of resources with functionalities of social networks. Users can save web resources within CROKODIL and annotate them by assigning tags. Specifically, the concept of semantic tagging [1] is applied, where a tag is associated with its semantic meaning by the selection of a tag type from a predetermined set. Learners can find, collect and organize web resources in semantic networks [12]. Semantic networks contain resources, tags and relations between them represented in a graphical notation consisting of nodes and edges.

2.2 Taxonomies in a collaborative E-Learning scenario

Taxonomies are commonly defined as a classification of concepts arranged in a hierarchical structure. This structure is typically organized by *is-a* relations. An *is-a* relation is a relation between a hypernym and a hyponym when the *hyponym* is a (kind of) *hypernym*. A taxonomy can be helpful in different aspects. We focus on the use of taxonomies as the basis for a recommender system in CROKODIL:

- Recommendation of tags: Tag recommendation mechanisms ease the process of finding good tags for a resource, but also consolidating the tag vocabulary. In figure 1 this is shown by the recommendation of the related tags "vehicle" and "limousine", where user A tags some resources with "car", which is a hypernym of "limousine" and a hyponym of "vehicle".

– Recommendation of general or specific resources of a topic: Learners need different types of resources in different stages of expertise. Novices need more general resources giving a broad overview of the topic, whereas experts need more specific information as they want to deepen their knowledge. This is shown in figure 1 as we recommend user A a resource about "vehicles" or "limousine" depending on the expert level of the user. In this example, we would tend to recommend "limousine" as user A has already two resources about "cars" which are more specific than resources about "vehicles".
– Recommendation of persons: User recommendation allows finding people working on similar topics or being an expert in a topic. Using CROKODIL's community functionality, a learner can contact other people and get help on a topic or to simply learn together. For user A in figure 1 it may be helpful to speak with users B or C in order to discuss or ask some issues about cars.

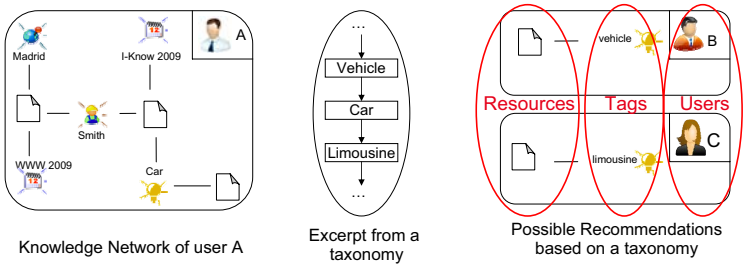

Fig. 1. Recommendations based on a Taxonomy

Based on the use cases that we want to support in CROKODIL, we identified domain-independency, actuality and multilingualism as relevant requirements for a taxonomy to support learners. Domain-independence in order to have support for a large coverage of topics, actuality, in order to process current topics and multilingualism in order to process information language independently. We use Wikipedia as data basis of our taxonomy as it satisfies all these requirements: It provides a good coverage offering more than 3'000.000 of concepts (English version), more than 1000 new articles per day and Wikipedia versions for 281 languages. However, most approaches in previous work have been developed especially for the English Wikipedia. These approaches provide good results and large coverage, but also incomplete socio-cultural knowledge for different languages. Socio-cultural knowledge is knowledge of cultural, regional or local relevance or importance. This kind of knowledge does not always exist in the English Wikipedia, e.g. the resistance of the German anti-nuclear movement to the "Castor-Transport", i.e. the transport of radioactive waste. In E-Learning scenarios, socio-cultural knowledge is crucial, for example, in school scenarios where students are doing homework about a regional topic or research groups in universities working on national projects.

In this paper, we present a method for a language-independent recognition of hyponymy relations towards taxonomy acquisition using Wikipedia. This method can be applied to a single Wikipedia version in order to acquire large-scale taxonomies with socio-cultural knowledge. We can cope with the multilingualism of the the resources in CROKODIL by using large-scale taxonomies in different languages depending on the languages used by the learners.

3 Related Work: Generating Multilingual Knowledge Bases from Wikipedia

Taxonomies and ontologies are often called knowledge bases as they provide information that can be used for knowledge derivation. We present in this section an overview of existing multilingual approaches and show that these approaches are not suitable for our application scenario. Wikipedia is a very popular source for the creation of knowledge bases as it provides a lot of information: It consists of articles that describe a specific concept. Articles are often linked to other languages using so called *interlanguage links* and belong to categories that group articles in related topics. Further, articles commonly link to other related articles by *wikilinks*. Both, the category structure as well as the link structure can be represented as a graph and because of this reason they are called *Category graph* and *Wikilink graph* respectively.

3.1 Manually created multilingual Knowledge Bases

A popular manually created ontology is WordNet [5]. It defines different semantic relations like hyponymy or meronymy. However, WordNet does not contain terminology of very specific domains and emerging topics (e.g. "iPad"), because new concepts have to be added to the ontology by a group of experts. There are different projects to develop similar knowledge resources for other languages. An extensive overview can be found on the web page of the Global WordNet association[1]. However, many of the projects are still in progress, have been abandoned or are not freely available.

3.2 Multilingual Knowledge Bases based on Wikipedia

Many approaches have been developed in recent years that derive semantic information from Wikipedia articles. Most of these approach produce very accurate knowledge bases. However, these approaches are difficult to apply in other languages as they rely on external corpora (e.g. [10]) or on other knowledge bases (e.g. [13]). For a more extensive explanation of these approaches, see [6]. There are only few approaches being applied to languages other than English. WikiTaxonomy, for instance, has been applied to German [3] and Japanese [14]. However,

[1] `http://www.globalwordnet.org/` - retrieved 01.06.2011

the authors of both papers state that they could not fully use the original approach as the articles in German and Japanese do not have the same coverage as in English.

Researchers have recognized the need for multi-lingual knowledge bases in other languages. The most straightforward method to develop multilingual knowledge bases is to use the interlanguage links [8]. However, interlanguage links do not exist for all articles. Because of this, Navigli et al. [7] construct a multilingual ontology using the Google Translate API[2] and a language corpus to get translations for non-existing interlingual links. However, socio-cultural knowledge is not incorporated as the core of the approach is still the English Wikipedia and the other ontologies are aligned to this core. In contrast to this approach, MENTA [4] weights the taxonomic information contained in 127 different Wikipedia versions in order to obtain a multilingual taxonomy. MENTA applies heuristics to decide if entities in different languages belong together or not. This approach seems to be very promising, but it contains a lot of irrelevant information for CROKODIL. For example, it contains information in many foreign languages that are not used in CROKODIL.

Our goal is the development of a language-independent approach for taxonomy acquisition. One of the challenges arises from the fact that each language has its own syntax and grammar. Our approach should be applicable in different languages without major changes to the algorithm and thus be language-independent. Our approach only use Wikipedia for this task, instead of using additional corpora or human effort which is different for each language. Using language-specific resources does not only hinder portability to other languages but also in some cases affects the quality and completeness of information within the knowledge base and makes the approach dependent on third parties.

4 Language-independent acquisition of Hierarchical Relations

In the following subsections we present the different features we use to recognize *is-a* relations from Wikipedia. The proposed method uses a set of 16 features. These features are described in this section.

4.1 Our features

We take pairs of categories (c_i and c_j) from the category graph and apply our features to each of these pairs, called *link*. The returned values are used to build the *feature vector* of a link. This is shown in fig 2.

The feature vector is used by a machine learning classifier to determine if there is an *is-a* relation between both categories. In table 1, we present an overview of the used features.

[2] `http://code.google.com/apis/language/` - retrieved 01.06.2011

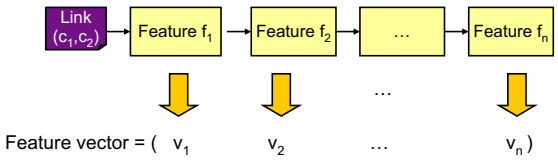

Fig. 2. Recommendations based on a Taxonomy

Preprocessing Features These features are used to filter out links containing nodes to administrative or refinement categories. Their value is `true`, if one of the given categories is an administration or refinement category. Administration categories contain prefixes like `Wikipedia` or `user`. Refinement links are used in Wikipedia to organize multiple categories using the pattern `X by Y` (e.g. "Companies by country"). The prefixes of administration categories and the preposition are adapted to the respective language. For instance, the prefix `User` is translated to `Benutzer` in German.

Table 1. Overview of the used features

Nr.	Name	Value type	Feature type
1	`adminCatFeature`	binary	preprocessing
2	`refinementLinkFeature`	binary	preprocessing
3	`positionOfHeadFeature`	$\{2, 1, 0, -1\}$	syntactic
4	`coocurrenceOfWordsFeature`	$\mathbb{N}$	syntactic
5	`coocurrenceFeature`	$\mathbb{N}$	structural
6	`c1NumberOfSubcategoriesFeature`	$\mathbb{N}$	structural
7	`c1NumberOfSuperCategoriesFeature`	$\mathbb{N}$	structural
8	`c2NumberOfSubcategoriesFeature`	$\mathbb{N}$	structural
9	`c2NumberOfSuperCategoriesFeatur`	$\mathbb{N}$	structural
10	`c1c2IncommingLinksFeature`	$\mathbb{N}$	structural
11	`c1c2OutgoingLinksLinksFeature`	$\mathbb{N}$	structural
12	`CommonWikilinksFeature`	$\mathbb{N}$	structural
13	`firstSentenceFeature`	binary	article
14	`c2Inc1Feature`	$\mathbb{N}$	article
15	`c1ArticleFeature`	binary	article
16	`c2ArticleFeature`	binary	article

Syntactic Features Syntactic features use string matching of syntactic components. `positionOfHeadFeature` uses the fact that the lexical head of two category names is a very effective method for labeling *is-a* links [10]. This feature returns a value $[2, -1]$ for pairs of categories c_1 and c_2 representing the position of the lexical head of the superordinate category. We differentiate between the following cases:

$$f_3(c_1, c_2) = \begin{cases} 2 & \text{if lexical head of } c_2 \text{ is at the end of } c_1 \\ 1 & \text{if lexical head of } c_2 \text{ is in the middle of } c_1 \\ 0 & \text{if lexical head of } c_2 \text{ is at the beginning of } c_1 \\ -1 & \text{else, i.e. no occurrence} \end{cases}$$

For instance, the value of this feature for c_1 ="French Revolution" and c_2 = "Revolution" is 2. The lexical head of category's lemma in English is usually the last word. This heuristic works for other languages as well: In Arabic, for instance, where the lexical head is at the beginning of a category name or in German, where the lexical head is "hidden' inside noun compounds where the multiple noun modifies the meaning given by the last one, e.g. "Baumhaus" (Eng. tree house). For languages like German, we "simulate" this heuristic using a constant matching window by taking the last x characters of c_2's name as the lexical head. The optimal value of x can be determined empirically by trying different window lengths. If the position of the head is not the head position in a given language then this is a sign that there is a *not-is-a* relation between c_1 and c_2. `cooccurrenceOfWords` represents cooccurrences of words in both category names.

Structural Features These features exploit the structure of the category graph and the wikilink graph. `coocurrenceFeature` returns `true` for pairs of categories which have at least one article in common [10]. `c1c2IncommingLinksFeature` and `c1c2OutgoingLinksFeature` count the number of articles in c_1, which have at least one (incoming or outgoing) link to any article in c_2. Their goal is to measure the strength of the relation between both categories [2]. `c1NumberOfSubcategories`, `c1NumberOfSupercategories`, `c2NumberOfSubcategories` and `c2NumberOfSupercategories` just count the number of sub- and supercategories of c_1 and c_2. Finally, `CommonWikilinksFeature` counts the number of common wikilinks between c_1 and c_2.

Article Features This set of features is applied to the content of articles. The first sentence of an article has a special meaning for taxonomic applications as it usually contains a definition of the concept [9]. This fact is used by `definitionSentenceFeature` to recognize *is-a* relations not in the whole article, but in the first sentence. This means that if an article a belongs to a category c_1 with both having the same label, then we search for occurrences of lexical heads of c_2 in a in the first sentence of the article. For instance, if c_1 = "Mice" and c_2 = "Pet Rodent", we test if the first sentence of the article "Mice" contains the term "rodent". If the check is positive, then this feature returns `true`. An advantage of this method is that a language-dependent search of patterns is not needed, and thus it can be applied in different languages. Further, the feature `c2Inc1Feature` counts the number of occurrences of the lexical head of c_2 in the rest of the article of c_1. `c1ArticleFeature` and `c2ArticleFeature` match c_1 and c_2 to Wikipedia articles. If c_2 can be matched to an article then we assume that this category is an existing concept, otherwise it may be a category used to structure the category graph like lists do.

4.2 Language-independency of our Approach

The approach presented here is language-independent as it is applicable as such to very different languages without modifying the features. Specifically, it can

be used to derive taxonomies from different languages with little information about a language needed. The mandatory data needed to run this approach is a list of prefixes of meta-categories that Wikipedia uses in this language, a list of prepositions contained in refinement-links (e.g. "by") and a list of prepositions of a language in order to match heuristically the lexical head of categories containing prepositions, e.g. "Battalions of the Canadian Expeditionary Force". The language-independency of our approach is restricted by the 281 existing Wikipedia version (i.e. we can not acquire taxonomies from other languages) and by the input of the prefixes and the prepositions mentioned before. Using only this information it is possible to generate taxonomies in different Wikipedia languages as we show in the next section.

5 Evaluation

We evaluated our approach in four different languages: three European languages (English, German and Spanish) and one language with non-latin characters (Arabic). We used a manually labelled corpus for each language to obtain results by applying each feature separately and in combination with the rest of the features for multiple languages.

Our corpus consists of 1000 randomly selected Wikipedia articles and categories. We extracted the corpus using the Wikipedia's export page[3]. After we extracted the corpus, we labelled it manually with the relevant relations (*is-a* and *not-is-a*). For our evaluation we choose decision trees as a classifier. All classification results were subjected to ten-fold cross validation, meaning that the corpus is partitioned in ten sub-samples with nine of these being used as training data and the last sub-sample being used as test data. Table 2 gives an overview of our results. It shows correctly and incorrectly classified instances. On average, 82.6 % of the labelled links are labelled correctly and 17.5 % are labelled incorrectly.

Table 2. Summarized results of our approach by languages

Language	English	Spanish	German	Arabic
Correctly classified instances	3539 (81.4 %)	1841 (84.7 %)	1968 (81.8 %)	2255 (82.3 %)
Incorrectly classified instances	806 (18.6 %)	333 (15.3 %)	437 (18.2 %)	484 (17.7 %)
Total number of instances	4345	2174	2405	2739

In Table 3, we take an in-depth look at the obtained results. It summarizes the most common metrics of evaluation of categorization algorithms: Precision, Recall and F-Measure. For Precision, we obtained average results of 76.1 % and Recall was 73.2 % for *is-a* relations and for *not-is-a* relations Precision was 81.2% and Recall 87.8 %. In order to understand the obtained values, we can take a look at the confusion matrix for English in Table 4.

[3] `http://en.wikipedia.org/wiki/Special:Export` - retrieved 01.06.2011

Table 3. Detailed accuracy by class and language

	Precision	Recall	F-Measure	Class
English	71 %	63.9 %	67.3 %	is-a
	85.3 %	88.9 %	87.1 %	not-is-a
Spanish	80.8 %	75.6 %	78.1 %	is-a
	86.7 %	89.8 %	88.2 %	not-is-a
German	74.1 %	70.5 %	72.3 %	is-a
	85.5 %	87.5 %	86.5 %	not-is-a
Arabic	78.8 %	78.5 %	78.6 %	is-a
	84.8 %	85.0 %	84.9 %	not-is-a

Table 4. Confusion matrix for English

a	b	← classified as
829	468	a = is-a
338	2710	b = not-is-a

The confusion matrix shows that the major sources of misclassification are incorrectly classified *is-a* links. The reason is that a high number of *is-a* instances could not be matched by any feature. This led to a misclassification of these instances. One possibility to improve the results presented here is to use interlanguage links to integrate the results from different languages or to improve single features, e.g. `firstSentenceFeature` may be improved by not only matching the category name, but also synonyms contained in redirect pages.

6 Conclusion and Future Work

In this paper, we presented a scenario of resource-based learning using web resources. We described our research prototype CROKODIL that aims to support self-directed learning and describe possible recommendations based on a taxonomy. We analyzed related work on the basis of the requirements of our scenario, identifying the need for language-independent extraction of hierarchical relations from Wikipedia in order to build a large-scale, up-to-date taxonomy in different languages. We described a robust language-independent approach which does not depend on other external sources of knowledge. Eventually, we evaluated the proposed features by measuring the accuracy of the classification of instances for each language. A direct comparison of the results with other knowledge bases is not easily possible as each knowledge base was evaluated using different evaluation data. On the other hand, approaches based on additional knowledge bases (e.g. WordNet) usually perform better than generic approaches as they are optimized with language-specific methods to work in one specific language, e.g. English.

In future work, we will apply our approach to the category graph of the German and English Wikipedia. After that, we plan development of a recommendation system based on the extracted taxonomy. An open question is, whether and how learners benefit from the offering of general or more specific topics. Further, we see room for improvement of our language-independent approach: the use of semantic similarity of articles in order to recognize important links in articles and applying features over different languages are examples of proposed ideas in order to improve this approach. Finally, we want to refine the evaluation by focusing on the performance of single features in different languages. For instance,

it would be very interesting to know, why syntactic features performed better in English than in Arabic.

The research presented here provides a foundation on which further applications and research (e.g. in the field of Wikipedia Mining or attaching semantics to web resources) can be based. Generally, our approach enables us to automatically derive a taxonomy from Wikipedia for different languages using syntactical and structural features and reducing the dependency on third parties. This approach can also be used to create knowledge for evaluation of new ontologies or taxonomies in languages where manually created knowledge bases do not already exist.

Acknowledgments. This work was supported by funds from the German Federal Ministry of Education and Research under the mark 01 PF 08015 A and from the European Social Fund of the European Union (ESF). The responsibility for the contents of this publication lies with the authors.

References

1. Böhnstedt, D. et al.: Collaborative Semantic Tagging of Web Resources on the Basis of Individual Knowledge Networks. In Proceedings of 1st and 17th Int. Conf. on User Modeling, Adaptation, and Personalization UMAP (2009)
2. Chernov, S. et al.: Extracting Semantic Relationships between Wikipedia Categories. In 1st Int. Workshop: "SemWiki2006 - From Wiki to Semantic" (2006)
3. Kassner, L. et al.: Acquiring a Taxonomy from the German Wikipedia. In Proc. of the 6th Int. Conf. on Language Resources and Eval. (2008)
4. de Melo, G and Weikum, G.: MENTA: Inducing multilingual taxonomies from Wikipedia. In Prof. of the 19th Int. Conf. on Inf. and Knowledge Manag. (2010)
5. Miller, G.: WordNet: a lexical Database for English. Communications of the ACM Vol. 38, No. 11:39-41 (1995)
6. Mendelyan, O. et al.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67(9), 716-754 (2009)
7. Navigli, R. and Ponzetto, S. P.: BabelNet: Building a very large multilingual semantic network In Proc. of the 48th annual meeting of the ACL 2010 (2010).
8. Nastase, V. et al.: WikiNet: A very large scale multi-lingual concept network. In Proc. of the LREC 2010 (2010)
9. Nguyen, D.P.T. et al.: Subtree Mining for Relation Extraction from Wikipedia. In Proceedings of HLT-NAACL, pp. 125 - 128, (2007).
10. Ponzetto, S. P. and Strube, M.: Deriving a large scale taxonomy from Wikipedia. In Proc. of the Second Conference on Empirical Methods in NLP, pp- 117-124 (2007)
11. Anjorin, M. et al.: CROKODIL - a Plattform for Collaborative Resource-based Learning. To appear in Proc. of the 6th European Conference on Technology Enhanced Learning EC-TEL 2011 (2011).
12. Sowa, J.F.: Semantic Networks. In Shapiro, S.C., ed.: Encyclopedia of Artificial Intelligence. Vol. 2. (1992)
13. Wu, F. and Weld, D. S.: Open Information Extraction using Wikipedia. In Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (2010).
14. Yamada, L. et al.: Hypernym discovery based on distrib. sim. and hierarch. struct.. In Proc. of the 2009 Conf. on Empirical Methods in Natural Language (2009)