Christian Schönberg, Helmuth Pree, and Burkhard Freitag:
Rich Ontology Extraction and Wikipedia Expansion Using Language Resources.
In: Proc. of the 11th Int. Conf. on Web-Age Information Management (WAIM'10), Jiuzhaigou, China, 2010. LNCS, volume 6184.

Rich Ontology Extraction and Wikipedia Expansion Using Language Resources

Christian Schönberg*, Helmuth Pree, and Burkhard Freitag

University of Passau, Department of Informatics and Mathematics
94030 Passau, Germany
{Christian.Schoenberg, Burkhard.Freitag}@uni-passau.de,
pree@fim.uni-passau.de

Abstract. Existing social collaboration projects contain a host of conceptual knowledge, but are often only sparsely structured and hardly machine-accessible. Using the well known Wikipedia as a showcase, we propose new and improved techniques for extracting ontology data from the wiki category structure. Applications like information extraction, data classification, or consistency checking require ontologies of very high quality and with a high number of relationships. We improve upon existing approaches by finding a host of additional relevant relationships between ontology classes, leveraging multi-lingual relations between categories and semantic relations between terms.

Key words: Ontology extraction, Semantic web, Web ontology

1 Introduction

Background knowledge and standardised vocabularies in the form of ontologies have become important factors in current knowledge management applications. Information extraction tasks make use of controlled vocabularies to increase precision and recall by focussing their efforts on known relevant terms [1], which is especially important if the source documents are of poor quality [2]. Classification of documents requires background knowledge about terms of interest when constructing feature vectors [3].

In general, it is difficult to obtain ontologies of sufficient quality and range for complex tasks. Wikis with a community large enough to act both normative and statistically self-correcting like Wikipedia have become a source for harvesting ontologies [4, 5]. Knowledge that is implicitly modelled in a wiki is made explicit in the extracted ontology. Using the well known web ontology language OWL for representation, categories are mapped onto classes, articles onto objects, and various forms of references onto different relationships. Yet such wikis usually have a structure that is optimised for browsing, not for reasoning. They are also often incomplete in the sense that relevant relationships are missing in

* This work is partially funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under grant number FR 1021/7-2.

their structure. In this paper, we present a new approach to improve ontology extraction from wikis by enriching the ontology structure in two dimensions, and by increasing its structural cohesion.

The rest of this paper is organised as follows: Section 2 describes the problem addressed in this paper, Sect. 3 describes our approach to ontology extraction and expansion from wikis, while Sect. 4 shows evaluation results for this approach. Section 5 discusses related work, and Sect. 6 concludes this paper.

2 Problem Statement

As obtaining large ontologies of high quality is an expensive task, either in terms of work or in terms of money, one approach is to generate such ontologies automatically. But even ontologies generated from human-engineered sources like wikis lack the quality of human-created ontologies like WordNet [6] or OpenGALEN [7]. In order to reduce the gap between generated and manually created ontologies, we propose an approach to enhance ontologies generated from wikis.

A wiki $W_L = (C_L, s_L, r_L, A_L, m_L, n_L)$ in a language L, i.e., the English Wikipedia, consists of a set of categories C_L, a sub-category relation $s_L : C_L \times C_L$ that defines a graph of sub-category relationships on C_L, a reference relation $r_L : C_L \times C_L$ that defines a graph of non-sub-category relationships on C_L, a set of articles A_L, a category-membership relation $m_L : A_L \times C_L$ that defines to which categories any given article belongs, and a naming function n_L that assigns a unique name to each category $c \in C_L$ and to each article $a \in A_L$. All articles consist solely of a descriptive text, i.e., their wiki content. The graph defined by s_L and r_L is also called the *category graph*. A complete wiki $W = (\mathcal{L}, \mathcal{W}_\mathcal{L}, tc, ta)$, e.g., Wikipedia, consists of a set of languages $\mathcal{L}$, a set of language-dependent wiki instances $\mathcal{W}_\mathcal{L} = \{ W_L \mid L \in \mathcal{L}\}$, and two translation relations $tc : C_{L1} \times C_{L2}$ and $ta : A_{L1} \times A_{L2}$, where $L1 \neq L2 \in \mathcal{L}$, that define any known translations for both categories and articles. Note that category and article redirects have been omitted here for reasons of clarity.

Recall that in OWL an ontology $\mathbf{O} = (\mathbf{C}, \mathbf{subClassOf}, \mathbf{R}, \mathbf{I}, \mathbf{type}, \mathbf{P})$ consists of a set of classes $\mathbf{C}$, a relation $\mathbf{subClassOf} : \mathbf{C} \times \mathbf{C}$ that defines sub-class relationships on $\mathbf{C}$, a set of relations $\mathbf{R}$ on classes that define relationships between two classes that are not sub-class relationships, a set of individuals $\mathbf{I}$, a class-membership relation $\mathbf{type} : \mathbf{I} \times \mathbf{C}$ that defines the classes any individual belongs to, and a set of property relations $\mathbf{P}$ on individuals, where each $\mathbf{p} \in \mathbf{P}$ defines a property for an individual, e.g., $\mathbf{P} \ni \mathbf{birthdate} : \mathbf{I} \times \mathit{date}$.

The simplest way to generate an ontology $\mathbf{O}$ from a wiki W is to select a language L and to define a mapping ϵ_L from C_L onto $\mathbf{C}$, from A_L onto $\mathbf{I}$, from r_L onto $\mathbf{R}$, from s_L onto $\mathbf{subClassOf}$, from m_L onto $\mathbf{type}$, and from a set of attributes (e.g., names or birth dates) extracted from A_L onto $\mathbf{P}$. L is called *primary language* and determines the language of the names for classes and individuals. In the context of this paper, all names are unique identifiers. One of the contributions of this paper is a refinement of the two mappings from s_L and from r_L, which do not seem to have been considered before.

The simple mapping process described above can result in a highly fragmented ontology. Rather than relying on this "raw" ontology, one might strive for an ontology with missing relationships added and redundant classes eliminated. We propose to obtain this desired ontology by taking the following steps:

1. extract the category graph from the wiki,
2. expand the graph with additional references,
3. remove superfluous categories,
4. list articles and extract attributes (this is beyond the scope of this paper),
5. map the wiki structure to an ontology structure.

3 Extraction Approach

To test our approach we selected the well known online resource Wikipedia, using the complete sets of categories C_L and articles A_L for eleven languages: $\mathcal{L} = \{$Chinese, Dutch, English, French, German, Italian, Japanese, Russian, Serbian, Spanish, Ukrainian$\}$, with English as the primary language. With the help of language experts, a set of keywords was created for each language, such as indicators used by Wikipedia to separate category names from article names (e.g., "Category:" in English).

3.1 Extraction of Category Graph

Using the keywords provided by language experts, the categories along with their sub-categories, references, and translations are detected in the Wikipedia data for all languages in $\mathcal{L}$, thereby constructing a category graph for each language.

3.2 Expansion and Optimisation of Category Graph

With English as the primary language for the ontology generation, the category graph of the other ten languages is analysed to find similarities and additional relationships.

In the first step we look at the structure of different language versions and try to infer hidden structural relationships, such as sub-category relationships. Let c and d be two categories in the English category graph for which no structural relationship is known. Let t_c and t_d be their translations into some other language L. If t_c and t_d have a relationship in L we infer a new relationship of the same kind between c and d. Runtime performance for this process is enhanced by maintaining an indexed lookup table for the translations in main memory, one language at a time.

The second enrichment step is to regard semantic similarities between category names. To this end, a lexical database of the primary language is required (e.g., WordNet [6] for English). For each pair of categories, a score is calculated. If it exceeds a predefined threshold, a new reference between the two categories is created. The score value $s(n_1, n_2)$ for two category names n_1 and n_2 can be

calculated based on the frequency of terms appearing in both names or based on a known relationship between terms from both category names, derived from the lexical database:

$$s(n_1, n_2) = \frac{\sum_{i=1}^{l(n_1)} s_t(n_1[i], n_2)}{\frac{l(n_1)+l(n_2)}{2}}$$

where $n[i]$ denotes the i^{th} term occurring in name n, and $l(n)$ denotes the number of terms in name n. The score $s_t(t, n)$ of a term t w.r.t. a name n, and the semantic distance $s_d(t_1, t_2)$ between two terms are defined as:

$$s_t(t, n) = \sum_{i=1}^{l(n)} s_d(t, n[i]) \qquad s_d(t_1, t_2) = \begin{cases} 0.8 \text{ if } t_1 = t_2 \\ 0.5 \text{ if } t_1 \text{ related to } t_2 \\ 0.0 \text{ otherwise} \end{cases}$$

The threshold used to filter potential category relationships based on their score varies between different languages. To enhance runtime performance, the process employs an inverted index on all occurring terms. It is used to find pairs of category names with similar terms, so that a name does not have to be checked against all other names.

Categories whose purpose is simply to provide a more convenient way to display the information contained within are removed by merging them with a more general category that contains the same information. The two most prominent kinds of such categories are *order-by* categories and *list* categories. They are detected by means of keywords provided by language experts.

3.3 Creating the Ontology

After the category graph of W_{en} has been expanded with additional references and pruned of superfluous categories, W_{en} is mapped onto an ontology structure. Both sub-category and sub-class relationships are often used to define the structure of the category graph or the ontology graph, respectively. While the sub-category relationship is modelled as a sub-class relationship, it could also be expressed as the *narrower-than* relationship of the SKOS[1] vocabulary. The semantics of the sub-category relationship are broad enough to allow for both interpretations. The relationships introduced by the second enrichment step are modelled as SKOS *related* relationships.

4 Evaluation

In addition to using English as the primary language, the approach is shown to be transferrable to other languages by providing evaluation results for an alternative setting where German is selected as the primary language. German is the language of the second largest Wikipedia in terms of articles. Table 1 shows the number of relationships that were added to the original category graph in the two enrichment steps (*Exp1* and *Exp2*) and the number of categories and relationships that were removed in the third step (*Merge*) for both languages.

[1] SKOS: http://www.w3.org/2004/02/skos, last visited 04/2010

	Exp1 (en)	*Exp1* (de)	*Exp2* (en)	*Exp2* (de)	*Merge* (en)	*Merge* (de)
Categories	0	0	0	0	$-22,702$	$-1,278$
Relationships	$+146,912$	$+59,514$	$+144,318$	$+3,723$	$-92,773$	$-55,803$

Table 1. Changes in category and relationship numbers for English and German

The results show that the expansion based on language versions already provides substantial improvements over the original category references: an increase of 14 % in 178 *s* for the English version, and an increase of 46 % in 111 *s* for the German version, respectively. The latter rather large increment can be explained by the fact that *Exp1* for German could make use of the English Wikipedia, which is by far the largest language version and thus provides the highest absolute number of relationships to learn from.

The expansion based on lexical databases on the other hand requires much more time for the English version, but results in a similarly large addition of new category references in addition to the first step: an increase of 14 % in 15 *hrs*. For the German version, both the expansion time and the resulting number of references are considerably lower: an increase of 3 % in 8 *s*. The cause for this discrepancy is the lower quality and size of the German lexical database used[2]. However, the low results for *Exp2* are offset by the high results for *Exp1*. The merging step took 68 *s* and 3 *s* for English and German, respectively.

As there exists no predefined "complete" version of the ontology, no absolute recall value can be computed. However, a relative increase of 28 % and 49 % for English and German, respectively, show the effectiveness of the method. The good results for different primary languages also indicate the generality of the proposed methods.

For a qualitative evaluation, a number of independent testers were asked to rate the validity of the newly generated relationships in different random samples of 200 entries each. The resulting precision values lie between 96 % and 99 %, which is an indication of the general high quality of our results. Values still closer to 100 % are rarely encountered in ontology evaluation, because different people usually have different opinions on how an ontology should be modelled to represent the actual world more closely. Few factual errors in the Wikipedia content which slightly affect the precision values for the first expansion step are outside of our control and are hopefully corrected by the community.

5 Related Work

[5] propose extending the MediaWiki framework by adding semantic information to wiki pages during the process of authoring. Links and information carrying text fragments are enriched by annotating semantic data.

In [8] a knowledge base derived from Wikipedia is presented. Articles that use infobox templates are treated as individuals, and attributes contained in the

[2] OpenThesaurus: http://www.openthesaurus.de, last visited 04/2010

template are extracted as properties. Additional relationships between entities are detected by using different types of indicators like similarity of attributes, Google query results, and WordNet relationships.

YAGO uses an extended RDFS OWL model consisting of individuals, classes and properties [4]. It uses Wikipedia and WordNet as sources, modelling categories and WordNet synsets as classes, articles as individuals, and category and synset relationships as class relationships. Properties are derived from articles.

6 Conclusion

In this paper, we have presented an approach to ontology extraction from wikis that improves upon existing methods by enriching the ontology structure with relationships inferred from different language versions and from semantic similarities between OWL classes. By removing classes with no inherent meaning of their own, the reduced ontology provides a more concise knowledge base. An increase of almost 30 % more relevant relationships for English and just under 50 % for German has been achieved during an evaluation using Wikipedia in eleven languages – including English, German, and Chinese – and with two different primary languages. The evaluation shows that the approach is effective, of sufficient generality, and can indeed deliver high quality results.

References

1. Claire Nédellec, Adeline Nazarenko: Ontology and Information Extraction: a necessary symbiosis. In Paul Buitelaar, Philipp Cimiano, B.M., ed.: Ontology Design and Population. IOS Press (2005) 155–170
2. K. Marko and S. Schulz and U. Hahn: MorphoSaurus - Design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. In: Methods of Information in Medicine. Number 44(4) (2005) 537–545
3. Chang, Y.: Automatically constructing a domain ontology for document classification. In: Int. Conf. on Machine Learning and Cybernetics. Vol. 4. (2007)
4. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW '07: Proc. of the 16th intern. conference on World Wide Web, New York, NY, USA, ACM Press (2007) 697–706
5. Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. Journal of Web Semantics **5** (2007) 251–261
6. Miller, G.A.: WordNet – a lexical database for the English language. http://wordnet.princeton.edu/ (2006)
7. Rogers, J.E., Roberts, A., Solomon, W.D., van der Haring, E., Wroe, C.J., Zanstra, P.E., Rector, A.L.: GALEN ten years on: Tasks and supporting tools. In: Proceedings of MEDINFO 2001, Amsterdam, IOS Press (2001) 256–260
8. Wu, F., Weld, D.S.: Automatically refining the Wikipedia infobox ontology. In: WWW '08: Proceeding of the 17th international conference on World Wide Web, New York, NY, USA, ACM (2008) 635–644